LEARN ONLY WHAT HELPS: TASK-PRIORITIZED DISTILLATION FROM HETEROGENEOUS TEACHERS FOR 2D MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate molecular property prediction (MPP) can benefit from heterogeneous molecular representations that encode complementary structural and chemical information. Yet obtaining such representations often requires additional molecular views or computations that are costly or impractical for large-scale screening. Knowledge distillation (KD) provides a natural way to transfer such supervision to a 2D graph student during training while retaining efficient 2D-only inference. However, existing multi-teacher KD methods largely focus on how to combine supervision from multiple teachers, but overlook whether a given teacher signal helps the downstream objective at a given student update. We argue that this is particularly significant for heterogeneous molecular teachers: differences in gradient magnitude can cause some teachers to dominate the shared update, while directional disagreement can make even a properly scaled teacher interfere with task descent. To address this challenge, we propose First-Order Routing of Teacher Expertise (FORTE), which reframes heterogeneous distillation as a task-prioritized routing problem rather than a teacher aggregation problem. Instead of assigning supervision to all available teachers, FORTE uses the downstream task gradient to determine both whether a teacher update is admissible and how strongly it should influence the student. Compatible teachers are routed through and a null route abstains when none qualifies, yielding an auxiliary update that is both aligned with and bounded by the task gradient. Theoretically, we show that, under the stated assumptions, FORTE improves the convergence bounds for the supervised downstream objective over conventional teacher aggregation, yielding a smaller sufficient iteration budget. Experimental results over a wide range of benchmarks show consistent improvements in both classification and regression, demonstrating the effectiveness and generality of our proposed method.

1 INTRODUCTION

Molecular property prediction has become an important component of modern molecular discovery, allowing large collections of candidate molecules to be evaluated computationally before more expensive experimental studies. Recent advances in molecular representation learning have substantially improved predictive performance across a broad range of tasks (Zheng et al., 2025; Li et al., 2025; Feng et al., 2025; Hong et al., 2026; Suryanarayanan et al., 2026). As these models are increasingly used for large-scale screening, however, their practical value depends not only on predictive accuracy, but also on the information required to make each prediction.

This consideration creates a fundamental trade-off in molecular representation. Two-dimensional molecular graphs are readily available from molecular connectivity and can be processed efficiently by graph neural networks, making them particularly attractive for large-scale deployment (Gilmer et al., 2017; Xu et al., 2019). Recent studies show that knowledge learned from 3D denoising and electron-density representations can improve molecular representation learning (Cho et al., 2025; Xiang et al., 2025). However, obtaining accurate 3D conformations for every candidate can be computationally expensive at scale (Cho et al., 2025). A desirable model should consequently benefit from rich molecular information during training without making that information a requirement at deployment. This motivates a training-rich, inference-light setting in which additional molecular views improve the learner while the final predictor continues to operate on 2D graphs alone.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

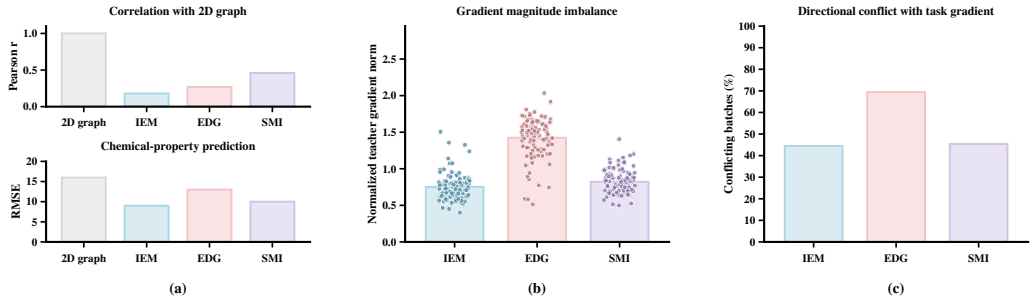

Figure 1: (a) Modality correlations and property-prediction errors, (b) average normalized gradient magnitudes, and (c) teacher–task conflicts on SIDER. The three teachers are the multi-view 3D image model IEM, the electron-density-aware structural-image model EDG, and the ChemBERTa-based SMILES sequence model SMI, whose gradients exhibit both scale imbalance and directional conflict.

Knowledge Distillation (KD) offers a direct way to realize this setting. A model trained on a richer molecular view can serve as a teacher, transferring useful knowledge to a 2D student that is retained at inference time (Hinton et al., 2015; Romero et al., 2015). The availability of several molecular views naturally suggests extending this idea to multiple teachers. At first glance, this seems advantageous: each teacher may contribute information that is absent from the others. The difficulty is that these sources of supervision ultimately act on the same student, and their influence on learning can be very different.

This difference is not merely a matter of assigning better weights. As illustrated in Figure 1, heterogeneous teachers can induce gradients with substantially different magnitudes, allowing one source of supervision to dominate the update regardless of its relevance to the downstream task. More importantly, even a properly scaled teacher can still push the student in a direction that conflicts with the task objective. This rules out some naive solutions for teacher aggregation. For example, equal weighting ignores both issues, while gradient-magnitude normalization addresses only the first (You et al., 2017; Chen et al., 2018; Yu et al., 2020). Thus, the central difficulty of heterogeneous distillation is not that teacher knowledge is scarce, but that available knowledge is not necessarily beneficial knowledge.

This observation calls for a different view of multi-teacher distillation. Rather than asking only how several teachers should be combined, we argue that one should first determine whether a teacher should be allowed to influence the student at all. This distinction also separates our setting from conventional multi-objective optimization (MOO). MOO methods designed to balance competing objectives typically seek a compromise among them (Sener & Koltun, 2018; Liu et al., 2021), whereas heterogeneous distillation has an inherently asymmetric goal: the supervised downstream objective defines what the student is ultimately expected to solve, and teacher objectives are useful only to the extent that they support it. Teacher supervision should therefore be treated as conditional assistance to the task, rather than as an objective that must be preserved for its own sake.

Following this principle, we propose First-Order Routing of Teacher Expertise (FORTE), a task-prioritized framework for heterogeneous molecular distillation. FORTE uses the downstream task gradient as the reference for determining teacher influence. It first calibrates teacher gradients so that their scale does not dictate the optimization, then suppresses supervision that conflicts with task descent. The remaining teachers are routed according to their task-relative contribution, while a null route allows the student to disregard auxiliary supervision when there is insufficient evidence that it is useful. In this way, FORTE does not assume that every available teacher should contribute. Instead, it makes teacher influence conditional on whether that supervision supports the downstream objective.

Beyond this routing principle, we provide a theoretical characterization of its optimization behavior. Under certain conditions, FORTE directly accelerates convergence of the target downstream task relative to conventional teacher aggregation. Experiments across 12 molecular property benchmarks and 3 graph backbones show consistent improvements in both classification and regression, and our routing analysis further confirms that FORTE suppresses task-conflicting teacher contributions while retaining useful auxiliary supervision.

We position our work relative to existing methods in Appendix A.

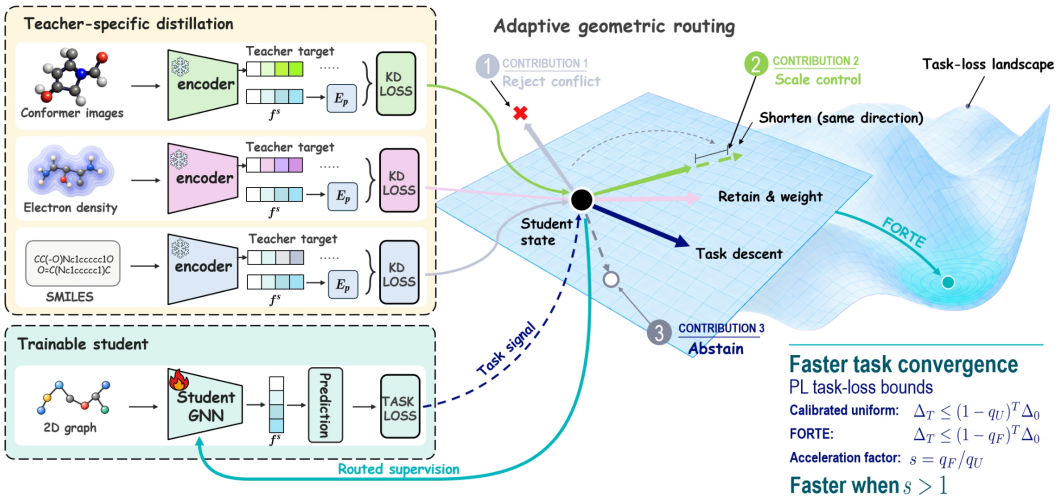

Figure 2: FORTE uses the task gradient as a geometric reference in student representation space, rescaling teacher gradients without rotating them and excluding directions outside the task-compatible half-space before combining the retained gradients. Under the stated assumptions, this geometry removes the residual error term of raw uniform weighting and yields a tighter convergence bound than normalized uniform weighting when the derived speedup factor exceeds one.

Our contributions are summarized as follows:

- We identify **task compatibility** as the central criterion for heterogeneous molecular distillation and formulate teacher selection as a task-prioritized routing problem.
- We introduce **FORTE**, which calibrates heterogeneous teacher supervision, suppresses task-conflicting contributions, and abstains when auxiliary evidence is insufficient, while preserving 2D-only inference.
- We establish theoretical guarantees for the optimization behavior of FORTE and demonstrate its effectiveness across **12 molecular property benchmarks and 3 graph backbones**.

2 First-Order Routing of Teacher Expertise (FORTE)

In this work, we transfer knowledge from informative but computationally expensive molecular modalities to an efficient 2D predictor. Models trained on rich modalities serve as teachers, while the student learns from their representations during training and uses only the 2D molecular graph at inference. The main challenge is that the task and multiple distillation losses update the same student encoder, and a teacher may be harmful when its gradient conflicts with the task gradient. We therefore propose *First-Order Routing of Teacher Expertise* (FORTE), which selects teachers based on the magnitude and direction of their first-order gradients. FORTE calibrates gradient magnitudes, routes only task-compatible teacher gradients, and includes a null option to ignore all teachers when none is useful. Under the stated geometric and smoothness assumptions, FORTE admits task-convergence bounds and explicit control of current-batch scale noise.

Figure 2 summarizes the training and deployment paths.

2.1 A Naive Method for Multi-Teacher Distillation

The Student Model. Because FORTE ultimately targets efficient inference from readily available molecular information, the student operates exclusively on 2D molecular graphs. For molecule i, let $\mathcal{G}_i$ denote its molecular graph and $\boldsymbol{y}_i$ its property label. The student comprises a graph encoder $f_{\boldsymbol{\theta}}$ and a task head $q_{\boldsymbol{\phi}}$. The encoder first maps the graph to a molecular embedding, after which the task head produces the property prediction:

$$\boldsymbol{h}_i = f_{\boldsymbol{\theta}}(\mathcal{G}_i) \in \mathbb{R}^d, \qquad \widehat{\boldsymbol{y}}_i = q_{\boldsymbol{\phi}}(\boldsymbol{h}_i), \qquad \ell_i^{\sup} = \ell_{\sup}(\widehat{\boldsymbol{y}}_i, \boldsymbol{y}_i). \tag{1}$$

Here, $\ell_{\sup}$ is the supervised prediction loss, and $\ell_i^{\sup}$ is its value for molecule i.

The Teacher Models. We index the teacher encoders by $\mathcal{K} = \{1, \ldots, K\}$, with $K = 3$ for the modalities introduced above. Each teacher is pre-trained on a distinct molecular modality and remains frozen during distillation. Given the modality-specific input $\boldsymbol{x}_{k,i}$, teacher k produces the embedding $\boldsymbol{z}_{k,i} = t_k(\boldsymbol{x}_{k,i}) \in \mathbb{R}^{d_k}$. Since the student and teacher encoders generally produce embeddings with different dimensions and representation spaces, their outputs cannot be aligned directly. FORTE therefore associates each teacher with a projection head $p_{\boldsymbol{\psi}_k} : \mathbb{R}^d \to \mathbb{R}^{d_k}$, where $\boldsymbol{\psi}_k$ denotes the trainable parameters of the projection head for teacher k. This head maps the student embedding into that teacher's representation space. The teacher-specific projection heads allow a single student encoder to receive supervision from multiple heterogeneous modalities.

A Naive Multi-Teacher Distillation Objective. Once the student embedding has been mapped into each teacher space, a natural baseline is to combine the supervised objective with representation-matching losses from all available teachers. For an observed pair of molecule i and teacher k, the distillation loss is defined as

$$\ell_{k,i}^{\mathrm{kd}} = \mathrm{dist}_k(p_{\boldsymbol{\psi}_k}(\boldsymbol{h}_i), \boldsymbol{z}_{k,i}), \tag{2}$$

where dist_k may be instantiated as mean squared error, cosine distance, or KL divergence, depending on the form of the teacher representation. In practice, heterogeneous molecular datasets often provide only a subset of modalities for each molecule. Consequently, some teacher embeddings are unavailable and should not contribute to the corresponding training objective. We record modality availability using $m_{k,i} \in \{0, 1\}$ and define the available set $\mathcal{A}_i = \{k \in \mathcal{K} : m_{k,i} = 1\}$, with $n_i = \max\{1, |\mathcal{A}_i|\}$ denoting the number of available teachers. Under this scenario, a naive way to combine the supervised and distillation objectives is to assign equal weight to every available teacher:

$$\ell_{\mathrm{uni},i} = \ell_i^{\mathrm{sup}} + \frac{\lambda}{n_i} \cdot \sum_{k \in \mathcal{A}_i} \ell_{k,i}^{\mathrm{kd}}. \tag{3}$$

Although uniform weighting provides a simple way to aggregate heterogeneous supervision, equal coefficients do not imply equal influence on optimization. The teachers having a conflict supervision signal with the task loss should be wisely ignored. This insight motivate a closer examination of the student-encoder update, which we develop in the next section.

2.2 The Sources of Multi-Teacher Conflict

To understand how uniform teacher weighting can hinder the downstream task, we examine the interaction between each distillation gradient and the supervised gradient. Although the uniform objective assigns equal loss weights to all available teachers, their actual influence is determined by the gradients applied to the shared student embedding. We therefore define the task gradient and the gradient induced by teacher k as

$$\boldsymbol{g}_i = \nabla_{\boldsymbol{h}_i} \ell_i^{\mathrm{sup}}, \qquad \boldsymbol{u}_{k,i} = \nabla_{\boldsymbol{h}_i} \ell_{k,i}^{\mathrm{kd}}. \tag{4}$$

Consider a small representation-space update involving teacher k, $\boldsymbol{h}_i^+ = \boldsymbol{h}_i - \eta \cdot (\boldsymbol{g}_i + \lambda \cdot \boldsymbol{u}_{k,i})$. A first-order expansion of the supervised loss gives

$$\ell_i^{\mathrm{sup}}(\boldsymbol{h}_i^+) - \ell_i^{\mathrm{sup}}(\boldsymbol{h}_i) = -\eta \cdot \|\boldsymbol{g}_i\|^2 - \eta \cdot \lambda \cdot \boldsymbol{g}_i^\top \boldsymbol{u}_{k,i} + O(\eta^2). \tag{5}$$

Equation (5) reveals two sources of conflict:

- **Magnitude Conflict.** The influence of teacher k depends on $\|\boldsymbol{u}_{k,i}\|$, since $|\boldsymbol{g}_i^\top \boldsymbol{u}_{k,i}| \leq \|\boldsymbol{g}_i\| \cdot \|\boldsymbol{u}_{k,i}\|$. Large differences in gradient norm can therefore allow a small number of teachers to dominate the student update.
- **Directional Conflict.** A positive inner product $\boldsymbol{g}_i^\top \boldsymbol{u}_{k,i} > 0$ reinforces the decrease induced by the supervised gradient, whereas a negative inner product offsets that decrease and may lead to negative transfer.

2.3 The Proposed Method

Equation (5) not only identifies the two sources of conflict, but also suggests the order in which they should be resolved. FORTE first prevents any teacher from dominating the update through its gradient magnitude, then removes directions that conflict with the task, and finally determines whether the remaining teacher guidance is sufficiently informative to use.

Estimate and normalize gradient magnitudes. Because teacher gradients are eventually combined in a shared update, their magnitudes must first be placed on a comparable scale. A norm computed

from a single mini-batch, however, can be noisy and biased toward the molecules observed in that batch. FORTE therefore maintains a bias-corrected exponential moving average (EMA) of the gradient norm for each teacher. Let n denote the number of valid mini-batch observations received by teacher k. Whenever modality k appears in the current mini-batch $\mathcal{B}$, the estimate is updated as

$$y_{k,n} = \frac{\sum_{i\in\mathcal{B}} m_{k,i}\cdot\delta_{k,i}}{\sum_{i\in\mathcal{B}} m_{k,i}}, \qquad v_{k,n} = \rho\cdot v_{k,n-1} + (1-\rho)\cdot y_{k,n}, \qquad \widehat{\nu}_{k,n} = \frac{v_{k,n}}{1-\rho^n}, \tag{6}$$

where $\delta_{k,i} = \|\boldsymbol{u}_{k,i}\|$, $y_{k,n}$ is the observed mini-batch mean, and $\rho \in [0,1)$ controls the adaptation rate. If teacher k is absent from a mini-batch, both its EMA state and its observation count remain unchanged.

Using the resulting scale estimate, FORTE calibrates each teacher gradient against the task gradient. Let $\zeta_i = \|\boldsymbol{g}_i\|$, and let $\varepsilon_c > 0$ be a small numerical constant that prevents division by zero. The calibration coefficient $c_{k,i}$ and the normalized teacher gradient are

$$c_{k,i} = \mathsf{StopGrad}\left(\min\left\{c_{\max}, \frac{\zeta_i}{\max\{\delta_{k,i}, \widehat{\nu}_{k,n}, \varepsilon_c\}}\right\}\right), \qquad \bar{\boldsymbol{u}}_{k,i} = c_{k,i}\cdot\boldsymbol{u}_{k,i}, \qquad q_{k,i} = \frac{\|\bar{\boldsymbol{u}}_{k,i}\|}{\zeta_i}, \tag{7}$$

with $q_{k,i} = 0$ when $\zeta_i = 0$. The quantity $q_{k,i}$ is thus the teacher's *task-relative magnitude*: it states how large the calibrated teacher gradient is relative to the task gradient, and the construction guarantees $0 \le q_{k,i} \le 1$, so no teacher can exceed the task gradient in magnitude. Detaching $c_{k,i}$ further ensures that $\nabla_{\boldsymbol{h}_i}\{c_{k,i}\cdot\ell^{\mathrm{kd}}_{k,i}\} = \bar{\boldsymbol{u}}_{k,i}$, making $\bar{\boldsymbol{u}}_{k,i}$ the gradient used in the final update.

Filter conflicts and compute routing scores. Magnitude calibration prevents a teacher from overwhelming the task update, but a well-scaled gradient can still point in a harmful direction. FORTE therefore evaluates the directional agreement between each normalized teacher gradient and the task gradient. Since $c_{k,i}$ is nonnegative, normalization preserves direction, allowing the cosine similarity to be computed from the original gradients:

$$s_{k,i} = \frac{\boldsymbol{g}_i^\top \boldsymbol{u}_{k,i}}{\zeta_i\cdot\delta_{k,i}}, \qquad b_{k,i} = m_{k,i}\cdot\mathbb{I}[s_{k,i} > \gamma], \qquad e_{k,i} = b_{k,i}\cdot q_{k,i}\cdot s_{k,i}, \tag{8}$$

where zero-norm pairs receive $s_{k,i} = 0$ and $\gamma \ge 0$ is the compatibility threshold. The binary eligibility mask $b_{k,i}$ equals one only when teacher k is both available for molecule i ($m_{k,i} = 1$) and sufficiently aligned with the task gradient ($s_{k,i} > \gamma$), and equals zero otherwise. Among the retained teachers, the routing score $e_{k,i}$ combines directional agreement $s_{k,i}$ with the calibrated gradient magnitude $q_{k,i}$. A teacher consequently receives strong evidence only when its guidance is both compatible with the task and substantial enough to affect the update.

Reserve routing mass for abstention. After gradient normalization and conflict filtering, FORTE must determine whether the retained teachers provide enough evidence to justify distillation. Normalizing only across the retained teachers would always assign them positive weight, even when all routing scores are weak. FORTE therefore performs a second normalization that includes a null route, which represents using no teacher for the current molecule:

$$\begin{aligned} &d_i = \varepsilon_\omega + \sum_{j\in\mathcal{K}} b_{j,i}\cdot\exp(e_{j,i}/\tau), \quad \omega_{0,i} = \frac{\varepsilon_\omega}{d_i}, \qquad \omega_{k,i} = \frac{b_{k,i}\cdot\exp(e_{k,i}/\tau)}{d_i}, \\ &r_i = \sum_{k\in\mathcal{K}} \omega_{k,i}\cdot e_{k,i}, \qquad \pi_{k,i} = r_i\cdot\omega_{k,i}, \quad \pi_{0,i} = 1 - \sum_{k\in\mathcal{K}}\pi_{k,i} = (1-r_i) + r_i\cdot\omega_{0,i}. \end{aligned} \tag{9}$$

Here, $\omega_{k,i}$ is the initial routing weight of teacher k, while $\omega_{0,i}$ is the weight assigned to the null route. The temperature τ controls how sharply FORTE distinguishes among teachers, and $\varepsilon_\omega > 0$ reserves a baseline score for abstention. The aggregate evidence $r_i \in [0,1]$ then determines how much of the initial teacher allocation is retained: strong and consistent evidence yields a larger r_i, whereas weak evidence shifts more weight to $\pi_{0,i}$. The final abstention weight $\pi_{0,i}$ therefore combines the uncommitted mass $1 - r_i$ with the null-route mass $r_i\cdot\omega_{0,i}$. If no teacher passes the conflict filter, then $\omega_{0,i} = 1$, $r_i = 0$, and $\pi_{0,i} = 1$. In that case, no teacher contributes to the update, allowing the student to learn entirely from the supervised objective and avoid negative transfer.

Having calibrated, filtered, and weighted the teacher gradients, FORTE forms the routed gradient $\boldsymbol{a}_i = \sum_k \pi_{k,i}\cdot\bar{\boldsymbol{u}}_{k,i}$. The final objective and its embedding gradient are

$$\ell_{\mathrm{FORTE},i} = \ell^{\mathrm{sup}}_i + \lambda\cdot\sum_k \mathsf{StopGrad}(\pi_{k,i}\cdot c_{k,i})\cdot\ell^{\mathrm{kd}}_{k,i}, \qquad \nabla_{\boldsymbol{h}_i}\ell_{\mathrm{FORTE},i} = \boldsymbol{g}_i + \lambda\cdot\boldsymbol{a}_i. \tag{10}$$

The construction directly yields the local certificate

$$\boldsymbol{g}_i^\top \boldsymbol{a}_i = r_i^2 \cdot \|\boldsymbol{g}_i\|^2, \qquad \|\boldsymbol{a}_i\| \le r_i \cdot \|\boldsymbol{g}_i\|, \qquad 0 \le r_i \le 1. \tag{11}$$

Thus, the routed teacher gradient remains aligned with the task gradient and cannot dominate its magnitude. These properties provide the foundation for the convergence analysis in Section 2.4.

2.4 Theoretical Guarantees

Restricted Isometry Assumption: FORTE routes teacher supervision using latent-space gradients, whereas training updates the GNN parameters. We connect these spaces through a restricted scaled isometry assumption (Appendix B.2, Definition B.2). Here, $\boldsymbol{a}_i$ is the routed auxiliary gradient defined in Eq. (10). For a batch B of size b, define $\boldsymbol{g}_B = b^{-1/2}(\boldsymbol{g}_i)_{i\in B}$, $\boldsymbol{a}_B = b^{-1/2}(\boldsymbol{a}_i)_{i\in B}$, and $\boldsymbol{J}_B = b^{-1/2}[D_{\boldsymbol{\theta}}\boldsymbol{h}_i]_{i\in B}$. The assumption requires the Jacobian to approximately preserve squared norms, up to a common scale, on the task–auxiliary subspace:

$$\alpha_B(1-\delta_B)\|\boldsymbol{v}\|^2 \le \|\boldsymbol{J}_B^\top \boldsymbol{v}\|^2 \le \alpha_B(1+\delta_B)\|\boldsymbol{v}\|^2, \qquad \boldsymbol{v} \in \operatorname{span}\{\boldsymbol{g}_B, \boldsymbol{a}_B\}, \tag{12}$$

where $\alpha_B > 0$ and $0 \le \delta_B < 1$. Lemma B.3 transfers the latent-space certificate to GNN parameter space, ensuring nonnegative auxiliary alignment when $R_B \ge \delta_B$, where $R_B^2 = \sum_{i\in B} r_i^2 \|\boldsymbol{g}_i\|^2 / \sum_{i\in B} \|\boldsymbol{g}_i\|^2$, with value zero for a zero denominator. Appendix B.5 establishes sufficient conditions for this assumption in GCN, GIN, and MPNN architectures.

We compare raw uniform weighting from Eq. (3), normalized uniform weighting that applies Eq. (7) before averaging, and FORTE from Eq. (10). Let b_{raw} bound the unnormalized teacher-gradient norm. Let $(\gamma_{\mathrm{n}}, \kappa_{\mathrm{n}})$ denote the conflict margin and relative gradient norm for normalized uniform weighting, and let $(\mu_{\mathrm{f}}, \rho_{\mathrm{f}})$ denote the corresponding descent and update-norm factors for FORTE.

Theorem 2.1 (Three-way convergence comparison, informal)**.** *Under smoothness and restricted isometry,*

$$\textit{raw uniform} : O(t^{-1}) + O(\lambda^2 \cdot b_{\mathrm{raw}}^2), \quad \textit{normalized uniform} : \ c_{\mathrm{n}}/t, \quad \textit{FORTE} : \ c_{\mathrm{f}}/t, \tag{13}$$

$$s_{\mathrm{f}\leftarrow\mathrm{n}} = \frac{c_{\mathrm{n}}}{c_{\mathrm{f}}} = \left(\frac{\mu_{\mathrm{f}}}{1-\lambda\cdot\gamma_{\mathrm{n}}}\right)^2 \cdot \left(\frac{1+\lambda\cdot\kappa_{\mathrm{n}}}{\rho_{\mathrm{f}}}\right)^2, \qquad \lambda\cdot\gamma_{\mathrm{n}} < 1. \tag{14}$$

Normalization removes the nonvanishing error term of raw uniform weighting by scaling each teacher gradient with the task gradient. Normalized uniform weighting still incurs the conflict penalty $1-\lambda\cdot\gamma_{\mathrm{n}}$, while FORTE removes conflicting directions. The first factor in Eq. (14) measures the task descent recovered by avoiding negative transfer. The second measures the smaller update norm obtained by routing teachers instead of averaging them uniformly. Therefore, $s_{\mathrm{f}\leftarrow\mathrm{n}} > 1$ gives a smaller sufficient iteration budget for FORTE.

Theorem 2.2 (EMA variance reduction, informal)**.** *Under finite teacher-scale variance and Lipschitz routing, the conditional update variances obtained using an EMA over observed teacher embeddings and using the current batch estimate satisfy $\sigma^2_{\mathrm{EMA},t} \le \sigma^2_{\mathrm{inst},t}$, with strict inequality whenever scale noise and routing sensitivity are nonzero.*

The EMA reduces the variance of stochastic updates, while the current mini-batch cosine similarities determine teacher compatibility. At inference, FORTE discards the teachers, projection heads, EMA states, and router, and uses only $\widehat{\boldsymbol{y}} = q_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathcal{G}))$.

3 Experiments

Datasets and evaluation protocol. In the evaluation stage, we use 8 molecular property prediction datasets from MoleculeNet, including BBBP, BACE, ClinTox, HIV, MUV, SIDER, Tox21, and ToxCast, with ROC-AUC as the evaluation metric. For broader evaluation, we also test the effectiveness of FORTE on 4 regression datasets (ESOL, Lipophilicity, Malaria, and CEP) with root mean square error (RMSE) as the metric. Following the standard protocol, each dataset is divided into training, validation, and test sets with a strict scaffold split at a ratio of 8:1:1; the test set is never used for hyperparameter, checkpoint, or ensemble-coefficient selection. Each model is trained for 100 epochs on both classification and regression datasets. Please see Appendices C.1, C.3, and C.5 for implementation and backbone details, comparison protocols, and pre-trained teacher weights, respectively.

3.1 RQ1: Downstream Effectiveness

RQ1

Does FORTE consistently improve molecular property prediction across classification and regression tasks and different student backbones, while improving downstream optimization?

Classification Results. We first evaluate FORTE on 8 molecular property prediction datasets with 3 student backbones (GIN, MolCLR, and Mole-BERT), and Table 1 reports the main results. We observe that the student backbones equipped with FORTE achieve state-of-the-art performance. The average margins over the strongest comparator are **5.2 percentage points for GIN (D&D), 6.1 for MolCLR (IEM), and 3.5 for Mole-BERT (D&D)**. The gains hold across all 3 pre-trained graph backbones, showing that teacher supervision remains useful even when the student already benefits from graph pre-training. We further evaluate FORTE on a wider range of drug-discovery tasks, including 4 regression benchmarks. Please see Appendix C.6 for the corresponding comparison results.

Table 1: ROC-AUC (%) performance of different methods on eight molecular property classification datasets under scaffold splitting; reproduced results are reported as mean ± standard deviation over three random seeds (0–2), where available. Best and second-best results, ranked before rounding, within each backbone block are bold and underlined, respectively.

Type	Method	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BBBP	BACE	Average
Student	**GIN**	76.4 ±0.3	63.8 ±0.2	60.5 ±0.4	66.3 ±1.6	74.2 ±0.7	77.5 ±0.8	66.2 ±3.9	79.9 ±1.5	70.6
Fusion	+Uniform-Fused	76.1 ±0.2	63.3 ±0.2	59.7 ±0.3	62.4 ±2.0	75.8 ±2.2	77.3 ±1.7	65.0 ±1.0	79.7 ±1.2	69.9
	+MoE-Fused	76.8 ±1.0	64.0 ±1.3	60.5 ±0.3	63.2 ±0.7	77.7 ±1.9	78.2 ±1.7	68.0 ±2.5	80.1 ±0.8	71.1
Loss weighting	+Uniform-KD	76.6 ±0.4	63.5 ±0.5	59.9 ±1.0	62.9 ±3.0	78.0 ±0.7	77.8 ±0.2	67.6 ±3.8	79.8 ±1.4	70.8
	+Learned-Gate	75.6 ±0.2	63.7 ±0.3	60.3 ±0.6	62.9 ±4.7	77.3 ±1.4	77.6 ±1.3	67.2 ±3.2	78.5 ±0.1	70.4
	+Confidence-Gate	76.6 ±0.6	63.6 ±0.2	60.9 ±0.7	61.4 ±1.1	76.1 ±2.5	77.4 ±1.4	66.0 ±1.7	79.9 ±0.8	70.2
Gradient optimization	+GradNorm	76.3 ±0.5	63.0 ±0.3	60.0 ±1.0	64.3 ±2.3	74.6 ±1.0	76.2 ±0.8	65.6 ±2.7	79.1 ±1.4	69.9
	+PCGrad	76.6 ±0.6	63.1 ±0.6	60.4 ±0.1	61.3 ±7.0	75.2 ±0.6	77.1 ±0.5	66.7 ±2.4	80.4 ±0.9	70.1
	+CAGrad	75.8 ±0.8	64.0 ±0.6	60.7 ±0.7	66.5 ±4.8	70.8 ±1.9	77.8 ±1.5	70.7 ±1.9	79.3 ±0.3	70.7
Molecular distillation	+CCMD	68.6 ±0.2	56.5 ±0.3	51.1 ±0.9	46.5 ±6.5	57.7 ±9.0	66.9 ±1.9	58.8 ±0.8	63.5 ±7.6	58.7
	+MolKD	72.5 ±0.7	61.7 ±0.5	59.1 ±0.8	61.5 ±3.7	77.7 ±1.2	75.9 ±2.0	64.4 ±3.9	**82.6** ±3.6	69.4
	+IEM	74.5 ±0.4	62.5 ±0.8	59.1 ±1.7	62.6 ±4.1	77.7 ±2.9	77.9 ±1.3	69.3 ±1.9	77.7 ±3.5	70.2
	+D&D	76.3 ±0.2	64.4 ±0.1	60.0 ±0.2	65.6 ±1.6	**81.0** ±1.1	77.4 ±1.2	64.6 ±2.9	80.7 ±2.0	71.3
Ours	**FORTE**	**79.4** ±0.5	**67.6** ±0.2	**66.1** ±0.4	**86.1** ±3.3	78.7 ±1.0	**79.3** ±3.1	**73.0** ±5.6	81.6 ±1.2	**76.5**
Student	**MolCLR**	75.2 ±0.4	63.7 ±0.5	60.5 ±1.5	64.6 ±11.0	68.3 ±0.9	74.0 ±2.9	64.5 ±4.8	74.8 ±1.6	68.2
Fusion	+Uniform-Fused	74.7 ±0.2	63.2 ±1.0	61.7 ±1.7	63.8 ±3.9	74.2 ±2.3	77.0 ±0.4	66.8 ±4.0	75.7 ±4.2	69.6
	+MoE-Fused	75.0 ±0.3	63.2 ±0.7	60.6 ±0.9	57.4 ±2.7	74.7 ±2.3	74.7 ±2.0	68.5 ±2.6	71.4 ±4.6	68.2
Loss weighting	+Uniform-KD	74.7 ±0.2	63.3 ±0.8	60.8 ±1.2	60.8 ±7.2	74.0 ±0.4	74.8 ±2.1	67.9 ±3.5	75.8 ±2.3	69.0
	+Learned-Gate	74.3 ±0.1	62.5 ±0.5	62.1 ±1.6	61.0 ±3.1	71.9 ±2.8	78.2 ±0.3	68.3 ±3.3	76.1 ±1.2	69.3
	+Confidence-Gate	74.7 ±0.5	63.9 ±0.9	60.4 ±1.6	60.8 ±5.8	75.9 ±0.9	74.7 ±0.1	67.0 ±5.5	71.6 ±8.5	68.6
Gradient optimization	+GradNorm	75.0 ±0.2	62.8 ±0.1	59.2 ±1.4	61.7 ±3.0	65.9 ±0.6	76.7 ±1.4	69.6 ±2.3	73.7 ±7.6	68.1
	+PCGrad	74.6 ±1.2	63.5 ±0.5	61.3 ±0.4	63.5 ±2.8	67.1 ±0.5	76.0 ±0.9	64.7 ±4.3	77.2 ±3.6	68.5
	+CAGrad	74.6 ±0.6	62.5 ±0.2	61.6 ±0.2	63.2 ±3.9	65.9 ±0.7	78.1 ±0.7	67.7 ±2.8	77.8 ±0.2	68.9
Molecular distillation	+CCMD	63.6 ±9.8	53.2 ±0.7	50.8 ±2.5	55.4 ±4.9	51.5 ±3.5	68.7 ±3.2	50.2 ±5.5	74.1 ±1.1	58.4
	+MolKD	72.9 ±0.1	62.1 ±0.6	56.9 ±1.6	66.7 ±7.2	76.7 ±0.8	75.1 ±0.7	67.9 ±0.8	75.6 ±5.3	69.2
	+IEM	75.1 ±0.7	64.2 ±0.3	61.1 ±0.3	66.1 ±7.3	74.4 ±2.6	74.0 ±0.6	69.0 ±1.7	75.3 ±2.1	69.9
	+D&D	75.2 ±0.6	64.4 ±0.1	60.7 ±0.2	63.0 ±6.0	**79.0** ±1.4	76.4 ±1.4	66.3 ±3.4	73.5 ±3.1	69.8
Ours	**FORTE**	**79.2** ±0.6	**68.1** ±0.3	**67.2** ±1.0	**84.7** ±3.5	77.1 ±2.1	**78.8** ±0.3	**72.8** ±3.9	**80.2** ±5.2	**76.0**
Student	**Mole-BERT**	77.5 ±0.8	65.5 ±0.1	63.9 ±1.1	73.6 ±3.0	78.3 ±2.1	77.2 ±1.1	66.5 ±0.9	81.9 ±0.8	73.1
Fusion	+Uniform-Fused	76.9 ±0.5	65.1 ±0.7	64.1 ±0.5	73.5 ±2.8	76.8 ±1.6	77.4 ±1.2	67.4 ±1.1	82.4 ±1.2	73.0
	+MoE-Fused	76.6 ±0.4	65.6 ±0.6	63.5 ±1.4	73.6 ±2.6	79.8 ±1.3	77.3 ±0.1	67.5 ±1.2	82.4 ±1.2	73.3
Loss weighting	+Uniform-KD	76.8 ±0.4	65.0 ±0.4	65.1 ±1.5	74.0 ±2.7	76.9 ±2.6	77.7 ±0.6	67.4 ±1.0	82.2 ±0.9	73.1
	+Learned-Gate	76.5 ±0.8	64.2 ±0.3	63.8 ±1.3	70.8 ±4.3	72.9 ±2.5	78.1 ±0.7	67.5 ±1.2	82.5 ±0.9	72.0
	+Confidence-Gate	76.8 ±0.6	65.2 ±0.6	65.2 ±1.5	73.8 ±3.7	79.2 ±1.8	77.5 ±1.7	67.7 ±1.3	82.2 ±0.7	73.4
Gradient optimization	+GradNorm	76.9 ±0.6	64.4 ±0.2	64.9 ±1.5	73.2 ±2.2	69.4 ±2.7	77.2 ±0.8	67.4 ±1.0	81.9 ±0.3	71.9
	+PCGrad	76.8 ±0.6	64.7 ±0.3	63.4 ±1.4	72.4 ±4.4	69.1 ±0.3	77.9 ±0.6	67.4 ±1.3	81.8 ±0.3	71.7
	+CAGrad	76.2 ±0.5	64.3 ±0.4	64.3 ±0.7	67.3 ±3.5	69.0 ±2.4	78.3 ±0.7	67.5 ±1.1	82.0 ±0.6	71.1
Molecular distillation	+CCMD	68.6 ±0.5	56.3 ±0.1	56.1 ±0.8	49.3 ±9.6	62.5 ±0.7	66.1 ±0.6	64.5 ±2.2	70.8 ±1.7	61.8
	+MolKD	75.5 ±0.5	63.2 ±0.0	63.4 ±0.9	74.5 ±4.0	76.7 ±2.0	77.4 ±1.9	66.7 ±1.8	80.8 ±0.9	72.3
	+IEM	77.3 ±0.1	65.6 ±0.5	63.6 ±0.5	73.4 ±3.6	79.0 ±1.2	76.4 ±1.3	67.7 ±1.3	81.5 ±0.7	73.1
	+D&D	77.7 ±0.2	65.1 ±0.6	64.2 ±1.0	75.0 ±3.4	**79.9** ±2.7	77.2 ±0.3	68.1 ±1.4	82.1 ±1.3	73.7
Ours	**FORTE**	**78.1** ±0.4	**68.0** ±0.4	**67.1** ±1.0	**86.6** ±5.5	79.4 ±1.2	**80.1** ±0.5	**72.8** ±0.1	**84.9** ±0.9	**77.1**

Downstream Optimization Behavior. To examine optimization behavior, we compare the first 100 epochs of FORTE, Uniform KD, GradNorm, PCGrad, and CAGrad. Figure 3 presents the supervised training-loss trajectories on 4 datasets, with the remaining 4 reported in Appendix C.7. FORTE attains the lowest final mean supervised training loss on all 4 displayed datasets. Together with the test ROC-AUC results in Table 1, these trajectories show improvements in both predictive performance and minimization of the downstream training objective. The lower final supervised loss indicates that FORTE makes more effective progress on the downstream training objective than uniform distillation and the gradient-based alternatives.

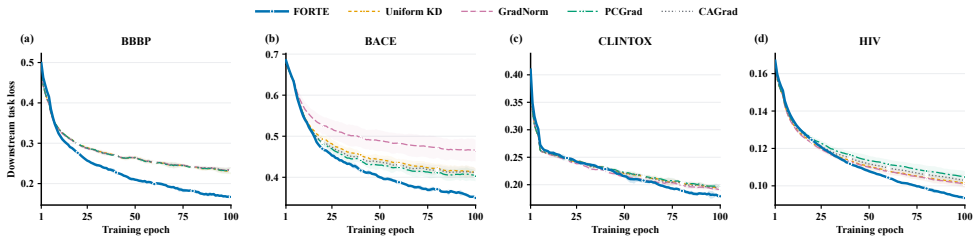

Figure 3: Supervised training loss on (a) BBBP, (b) BACE, (c) CLINTOX, and (d) HIV over 100 epochs; curves show 3-seed means with $\pm$ SD bands after 5-epoch smoothing.

Findings. FORTE improves prediction across all 3 pre-trained graph backbones and attains the lowest final mean supervised training loss on 7 of 8 datasets, supporting its effectiveness beyond a single architecture or task.

3.2 RQ2: Task-Conflict Resolution and Selective Routing

RQ2

Does FORTE selectively suppress teacher gradients that conflict with the downstream task while retaining positively aligned auxiliary supervision across molecular datasets?

To examine whether FORTE resolves directional conflicts rather than merely reducing the magnitude of auxiliary gradients, we analyze matched gradient updates at validation-selected checkpoints using GIN on SIDER, Tox21, BBBP, and BACE. Figure 4 compares the alignment of each matched update under equal weighting (x-axis) with its alignment after FORTE routing (y-axis). Under equal weighting, a subset of teacher contributions opposes the supervised gradient, as indicated by the negative-alignment region. FORTE largely suppresses these conflicting contributions while retaining positively aligned auxiliary updates, suggesting that routing selectively filters harmful supervision rather than discarding auxiliary knowledge altogether.

An extended analysis of these conflict sources and the resulting task-conditioned routing behavior across datasets is provided in Appendix C.9.

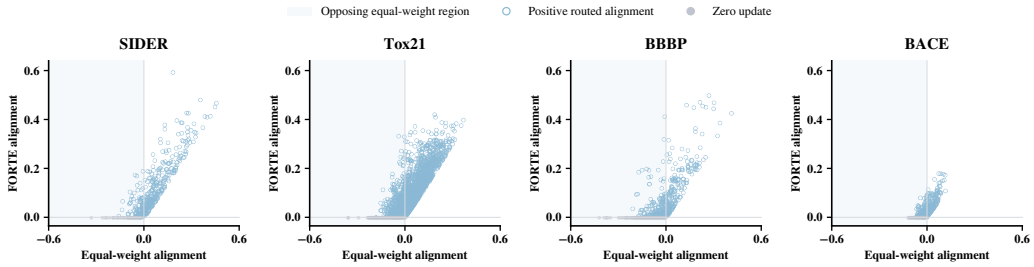

Figure 4: The x- and y-axes show normalized task-aligned auxiliary components under equal-weight mixing and FORTE routing, respectively, with each point comparing the same molecule at the same checkpoint in student representation space.

Findings. Our analysis reveals that FORTE consistently removes negatively aligned teacher contributions while retaining beneficial auxiliary updates, directly validating the intended conflict-resolution behavior.

3.3 RQ3: Routing Behavior and Ablation Study

RQ3

How does FORTE adapt heterogeneous teacher supervision across tasks, and how do its key routing components contribute to performance?

Figure 5 shows that FORTE assigns different proportions of auxiliary supervision to the 3 teachers across BBBP, BACE, ClinTox, and HIV. The allocations are generally consistent across seeds, although BACE exhibits greater variation. Results for the remaining datasets and additional routing diagnostics are provided in Appendix Figure 9 and Section C.9. To isolate the contribution of each routing component, we evaluate 3 ablations. We remove the teacher-to-task compatibility gate and cosine factor from the routing evidence (noted as *w/o Compatibility* in Tab. 2). We replace the historical norm reference with the current per-molecule teacher-gradient norm while retaining task-relative magnitude calibration (noted as *w/o EMA calibration* in Tab. 2). We normalize routing weights only over eligible teachers, assigning total mass 1 if any teacher is eligible and setting the auxiliary update to 0 otherwise (noted as *w/o Null* in Tab. 2). Table 2 shows that each ablation lowers average ROC-AUC across the 24 backbone–dataset combinations; removing compatibility filtering produces the largest decrease, while calibration and abstention provide additional gains.

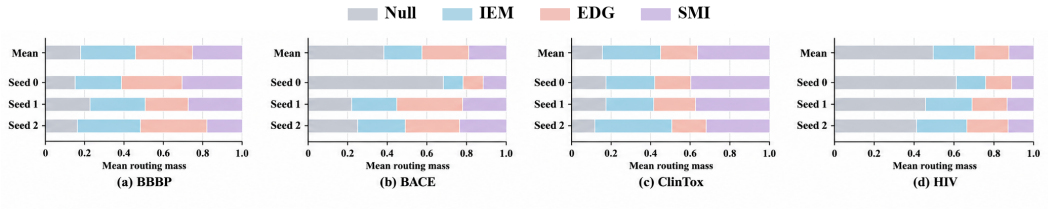

Figure 5: Mean and seed-wise routing masses assigned to Null, IEM, EDG, and SMI on (a) BBBP, (b) BACE, (c) ClinTox, and (d) HIV; gray segments denote the null route.

Table 2: Component ablations across 3 student backbones and 8 classification datasets, reported as mean ± standard deviation of ROC-AUC (%). Bold values denote the complete FORTE model.

Backbone	Method	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BBBP	BACE
GIN	**FORTE**	**79.4 ± 0.47**	**67.6 ± 0.24**	**66.1 ± 0.41**	**86.1 ± 3.28**	**78.7 ± 0.95**	**79.3 ± 3.06**	**73.0 ± 5.58**	**81.6 ± 1.17**
	w/o Null	78.62 ± 0.79	66.01 ± 1.98	65.39 ± 1.29	85.54 ± 2.94	78.05 ± 2.20	79.10 ± 1.86	70.75 ± 4.63	80.26 ± 1.05
	w/o EMA calibration	78.60 ± 0.46	66.16 ± 0.24	65.52 ± 0.67	84.10 ± 3.30	77.61 ± 2.40	78.49 ± 1.62	70.23 ± 3.69	79.83 ± 1.24
	w/o Compatibility	76.37 ± 0.69	64.32 ± 0.69	59.50 ± 0.46	68.54 ± 2.68	76.23 ± 2.09	75.29 ± 1.45	68.19 ± 4.48	80.54 ± 1.60
	Student only	76.43 ± 0.32	63.82 ± 0.18	60.52 ± 0.41	66.29 ± 1.61	74.23 ± 0.70	77.50 ± 0.78	66.18 ± 3.85	79.90 ± 1.50
MolCLR	**FORTE**	**79.2 ± 0.62**	**68.1 ± 0.30**	**67.2 ± 0.96**	**84.7 ± 3.54**	**77.1 ± 2.05**	**78.8 ± 0.32**	**72.8 ± 3.86**	**80.2 ± 5.24**
	w/o Null	78.23 ± 1.15	67.19 ± 0.40	66.51 ± 1.27	82.41 ± 2.61	76.79 ± 1.48	77.79 ± 0.27	71.99 ± 3.95	79.45 ± 1.85
	w/o EMA calibration	77.31 ± 0.39	67.01 ± 0.30	65.82 ± 1.06	81.27 ± 6.64	75.68 ± 1.45	77.68 ± 0.94	70.34 ± 2.94	78.60 ± 1.32
	w/o Compatibility	75.16 ± 0.57	63.70 ± 0.64	61.42 ± 0.49	69.13 ± 3.24	72.69 ± 2.52	76.86 ± 1.59	66.54 ± 3.12	75.53 ± 7.90
	Student only	75.15 ± 0.39	63.65 ± 0.49	60.45 ± 1.48	64.64 ± 11.01	68.34 ± 0.91	73.99 ± 2.86	64.54 ± 4.76	74.83 ± 1.57
Mole-BERT	**FORTE**	**78.1 ± 0.41**	**68.0 ± 0.39**	**67.1 ± 1.03**	**86.6 ± 5.50**	**79.4 ± 1.22**	**80.1 ± 0.50**	**72.8 ± 0.12**	**84.9 ± 0.91**
	w/o Null	77.16 ± 0.39	66.19 ± 0.21	65.09 ± 0.52	84.05 ± 5.32	78.77 ± 0.80	78.56 ± 1.24	71.77 ± 0.18	82.22 ± 1.35
	w/o EMA calibration	77.41 ± 0.53	67.17 ± 0.43	66.62 ± 1.10	85.16 ± 3.05	78.17 ± 0.95	77.67 ± 0.46	70.91 ± 0.16	81.08 ± 1.22
	w/o Compatibility	77.38 ± 0.07	65.18 ± 0.29	64.59 ± 1.14	75.17 ± 3.33	78.08 ± 0.64	77.26 ± 0.74	67.85 ± 0.04	81.66 ± 0.87
	Student only	77.51 ± 0.79	65.47 ± 0.14	63.91 ± 1.05	73.63 ± 2.97	78.31 ± 2.11	77.24 ± 1.14	66.51 ± 0.85	81.86 ± 0.78

Findings. FORTE adjusts teacher participation across tasks, and removing any of its 3 routing components consistently weakens predictive performance. Compatibility filtering contributes most, while gradient calibration and the option to abstain provide further gains.

4 Conclusion

We presented FORTE, a task-prioritized framework for heterogeneous molecular distillation that preserves 2D-only inference. By calibrating teacher gradients, suppressing task-conflicting supervision, and allowing abstention when auxiliary evidence is weak, FORTE produces task-aligned distillation updates. Experiments across 12 molecular property benchmarks and three graph backbones demonstrate consistent improvements, supporting task-conditioned teacher routing as an effective way to exploit heterogeneous molecular knowledge without increasing inference cost.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and LaTeX formatting. All technical content, mathematical derivations, claims, references, and final wording were reviewed and verified by the authors, who take full responsibility for the submission.

REFERENCES

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 794–803, 2018.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. In *Machine Learning for Molecules Workshop at NeurIPS*, 2020.

Sungjun Cho, Dae-Woong Jeong, Sung Moon Ko, Jinwoo Kim, Sehui Han, Seunghoon Hong, Honglak Lee, and Moontae Lee. 3D denoisers are good 2D teachers: Molecular pretraining via denoising and cross-modal distillation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, pp. 110–118, 2025.

Simon S. Du, Kangcheng Hou, Barnabás Póczos, Ruslan Salakhutdinov, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In *Advances in Neural Information Processing Systems*, pp. 5723–5733, 2019.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

Minggao Feng, Chengxi Zhao, Graeme M. Day, Xenophon Evangelopoulos, and Andrew I. Cooper. A universal foundation model for transfer learning in molecular crystals. *Chemical Science*, 16 (28):12844–12859, 2025.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Huiyang Hong, Xinkai Wu, Hongyu Sun, Chaoyang Xie, Qi Wang, and Yuquan Li. A hierarchical interaction message net for accurate molecular property prediction. *Communications Chemistry*, 9 (1):150, 2026.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17:1–13, 2012.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020a.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. GPT-GNN: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1857–1867, 2020b.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Biaoshun Li, Mujie Lin, Tiegen Chen, and Ling Wang. FG-BERT: A generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings in Bioinformatics*, 24(6):bbad398, 2023.

Na Li, Jianbo Qiao, Fei Gao, Yanling Wang, Hua Shi, Zilong Zhang, Feifei Cui, Lichao Zhang, and Leyi Wei. GICL: A cross-modal drug property prediction framework based on knowledge enhancement of large language models. *Journal of Chemical Information and Modeling*, 65(11): 5518–5527, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 18878–18890, 2021.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations*, 2022.

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 21497–21526, 2023.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, pp. 12559–12571, 2020.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pp. 991–1001, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pp. 527–538, 2018.

Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*, 2020.

Parthasarathy Suryanarayanan, Yunguang Qiu, Shreyans Sethi, Diwakar Mahajan, Hongyang Li, Yuxin Yang, Elif Eyigoz, Aldo Guzmán-Sáenz, Daniel E. Platt, Timothy H. Rumbell, Kenney Ng, Sanjoy Dey, Myson Burch, Bum Chul Kwon, Pablo Meyer, Feixiong Cheng, Jianying Hu, and Joseph A. Morrone. Multi-view biomedical foundation models for molecule-target and property prediction. *Advanced Science*, 13(14):e17840, 2026.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *International Conference on Learning Representations*, 2023.

Hongxin Xiang, Shuting Jin, Jun Xia, Man Zhou, Jianmin Wang, Li Zeng, and Xiangxiang Zeng. An image-enhanced molecular graph representation learning framework. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6107–6115, 2024.

Hongxin Xiang, Jun Xia, Xin Jin, Wenjie Du, Li Zeng, and Xiangxiang Zeng. Electron density-enhanced molecular geometry learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7840–7848, 2025.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised graph-level representation learning with local and global structure. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11548–11558, 2021.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, 2017.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal molecular pretraining via modality blending. In *International Conference on Learning Representations*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 5824–5836, 2020.

Liang Zeng, Lanqing Li, and Jian Li. MolKD: Distilling cross-modal knowledge in chemical reactions for molecular property prediction. *arXiv preprint arXiv:2305.01912*, 2023.

Hao Zhang, Nan Zhang, Ruixin Zhang, Lei Shen, Yingyi Zhang, and Meng Liu. Coordinating cross-modal distillation for molecular property prediction. *arXiv preprint arXiv:2211.16712*, 2022.

Xinyi Zhang, Yanni Xu, Changzhi Jiang, Lian Shen, and Xiangrong Liu. MoleMCL: A multi-level contrastive learning framework for molecular pre-training. *Bioinformatics*, 40(4):btae164, 2024.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 7(3):437–447, 2025.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *International Conference on Learning Representations*, 2023.

Appendix

Table of Contents

A Related Work 14
- A.1 Molecular Property Prediction with Diverse Molecular Representations 14
- A.2 Learning from Multiple Molecular Views 14
- A.3 Multi-Teacher Distillation under Conflicting Supervision 15
- A.4 Positioning FORTE 16

B Full-Data and Mini-Batch Parameter-Space Analysis 17
- B.1 Empirical objective, full-data direction, and actual mini-batch update 17
- B.2 The correct batch-level restricted-isometry event 18
- B.3 Full-data convergence after the one-time isometry transfer 20
- B.4 Full-data rate comparison with uniform and learned teacher weighting 21
  - B.4.1 Distillation-accelerated convergence of the supervised task loss 25
- B.5 GNN regimes in which the batch isometry holds 28
  - B.5.1 Concrete GCN, GIN, and multilayer message-passing students 28
- B.6 Actual mini-batch direction and high-probability convergence 38
  - B.6.1 Mini-batch task-loss contraction with direction error 42
  - B.6.2 High-probability comparison of stochastic orders 43
- B.7 Why the stochastic assumptions hold for bounded finite-depth GNNs 45

C Additional Experimental Material 47
- C.1 Implementation Details 47
- C.2 Datasets of Downstream Tasks 47
- C.3 Competitors 48
- C.4 Comparison with More State-of-the-Art Methods 50
- C.5 Teacher Sources and Feature Extraction 50
- C.6 Regression Results 51
- C.7 Additional Task-Loss Trajectories 53
- C.8 Teacher Availability and Routing Across Classification Tasks 53
- C.9 Extended Analysis of Task-Conditioned Routing 54

A Related Work

Our work is motivated by a practical tension in molecular property prediction: richer molecular representations can provide information beyond a 2D graph alone, but obtaining these representations can substantially increase data acquisition and computational costs. This has led to growing interest in how multimodal molecular information can be used in practical downstream tasks. We review this literature from three perspectives: the evolution of molecular representations for property prediction, learning from multiple molecular views, and knowledge distillation under conflicting auxiliary supervision.

A.1 Molecular Property Prediction with Diverse Molecular Representations

Molecular property prediction critically depends on how molecular structure is represented. Early graph-based approaches model molecules as 2D graphs, with atoms as nodes and chemical bonds as edges, and learn molecular representations through neighborhood message passing (Gilmer et al., 2017). More expressive graph neural networks further improve the ability to distinguish molecular structures and have become widely adopted backbones for molecular property prediction (Xu et al., 2019). Because molecular connectivity is readily available and inexpensive to process, 2D graph models provide an attractive balance between predictive capability and computational efficiency for large-scale molecular screening. Self-supervised molecular pre-training further strengthens these representations by learning transferable structural regularities from unlabeled molecules (Wang et al., 2022; Xia et al., 2023).

Nevertheless, 2D connectivity provides only one view of a molecule. Many molecular properties are affected by geometric and physicochemical factors that are not fully specified by a topological graph. Geometry-aware models therefore incorporate interatomic distances, angular information, and directional interactions to represent three-dimensional molecular structure (Schütt et al., 2017; Gasteiger et al., 2020). Large-scale 3D pre-training frameworks subsequently exploit geometric information to learn transferable molecular representations (Liu et al., 2022; Zhou et al., 2023). Other methods use molecular images or sequence-based representations to expose structural and chemical signals that differ from those encoded by a 2D graph (Xiang et al., 2024; Zheng et al., 2025; Li et al., 2025). These approaches show that changing the molecular view can change the information available to the predictor.

The resulting progression from 2D topology to geometry, images, and molecular language suggests that no single representation universally captures all information useful for molecular property prediction. At the same time, richer representations are not equally suitable for deployment. Obtaining specialized molecular views and processing them with their corresponding models can be substantially more expensive than inference from molecular connectivity alone. This motivates a training-rich, inference-light setting in which additional molecular representations contribute supervision during learning, while the deployed predictor retains the efficiency and availability advantages of a 2D graph model. FORTE operates in this setting, treating diverse molecular representations as potentially useful sources of training-time knowledge rather than mandatory inference inputs.

A.2 Learning from Multiple Molecular Views

The complementarity of molecular representations has motivated a parallel line of work on learning jointly from multiple molecular views. A prominent example is 2D–3D molecular representation learning, where topological and geometric information describe the same molecule from different perspectives. Existing approaches commonly employ separate encoders for different representations and encourage consistency through contrastive or predictive objectives. Such methods establish that cross-view information can enrich molecular representations beyond what is obtainable from either representation in isolation (Liu et al., 2022; Zhou et al., 2023).

Recent molecular representation methods also move beyond a single global embedding. They combine structural and visual signals at different levels of abstraction (Xiang et al., 2024). They also integrate molecular and semantic information to enrich prediction (Li et al., 2025; Suryanarayanan et al., 2026). This line of work asks how complementary molecular views should be represented and fused so that the learned representation becomes more informative. It provides strong motivation for exploiting heterogeneous molecular information during learning.

Our setting addresses a different issue that emerges once heterogeneous views are used as supervision for a common downstream predictor. FORTE does not require multiple molecular views to remain available to the deployed model, nor does it aim to construct a single fused multimodal representation. Instead, models specialized on rich molecular views remain frozen teachers during distillation, whereas the final student operates only on a 2D molecular graph. Under this training configuration, the relevant problem changes: even when several views are individually informative, their induced supervisory signals need not be mutually compatible with the property objective being optimized by the student. Different teachers can produce gradients with substantially different scales, and their directions can disagree with the supervised task or with one another. Thus, the existence of complementary molecular information does not by itself determine how that information should influence a shared student update. Resolving this optimization conflict is the focus of FORTE.

A.3 Multi-Teacher Distillation under Conflicting Supervision

Knowledge distillation provides a natural mechanism for transferring information from richer models into a lightweight predictor. Classical knowledge distillation trains a student to reproduce softened teacher predictions (Hinton et al., 2015), while representation-level distillation extends supervision to intermediate features learned by the teacher (Romero et al., 2015). With several teachers, multi-teacher distillation further combines knowledge from multiple networks, typically through averaging or other weighting mechanisms (You et al., 2017). These approaches establish the basic paradigm relevant to our setting: information accessible to expensive teachers during training can be transferred into a student that remains inexpensive at inference.

The presence of heterogeneous teachers, however, introduces an optimization problem that is not resolved by teacher aggregation alone. Conventional weighting schemes primarily determine how much each teacher contributes. Yet equal loss coefficients do not imply equal effects on the student because gradients generated by different teacher losses can differ substantially in magnitude. Adaptive weighting or normalization can mitigate this imbalance, but magnitude is only one source of interference. Even after rescaling, an auxiliary gradient can oppose the supervised task gradient and degrade the descent direction of the objective used to evaluate the student. A weighting mechanism that normalizes across the available teachers also assumes that some teacher contribution should always be retained, whereas heterogeneous supervision may require rejecting all auxiliary signals for a particular update.

These issues connect heterogeneous distillation to gradient-aware optimization. GradNorm adapts objective weights according to relative gradient magnitudes and training rates (Chen et al., 2018), while PCGrad and CAGrad address directional conflict in multi-task updates (Yu et al., 2020; Liu et al., 2021). Pareto-based formulations instead seek a compromise among objectives (Sener & Koltun, 2018). These methods show that interactions among learning signals should be considered at the gradient level rather than only through scalar loss weights. Their usual formulation is symmetric, however, whereas heterogeneous distillation has an asymmetric endpoint: the supervised property loss defines success and teacher matching is useful only when it supports that loss.

Prior gradient-aware auxiliary-learning methods, such as Du et al. (2018), use task-gradient similarity to determine whether auxiliary supervision should be retained. But it is insufficient for heterogeneous distillation in drug discovery. In realistic molecular prediction settings, this limitation arises from two coupled sources: heterogeneous molecular modalities induce teacher gradients with substantially different magnitudes, and even a properly scaled teacher can still produce a direction that conflicts with the downstream task gradient. These effects make simple gradient-similarity filtering inadequate for coordinating multiple heterogeneous molecular teachers. FORTE is therefore designed specifically for this setting. Our result addresses a strictly different—and stronger—question: whether auxiliary distillation can accelerate optimization of the main task itself in drug discovery. Du et al. (2018) establish that gradient-aware auxiliary learning can preserve convergence of the main task, but do not show that auxiliary supervision accelerates convergence of the main-task loss. In contrast, although FORTE optimizes a joint supervised–distillation objective, our analysis shows that, under the stated conditions, routed distillation yields a strictly stronger contraction of the supervised task loss itself than supervised gradient descent at the same admissible step size, together with a smaller sufficient iteration budget. Theorem B.16 answers this question affirmatively under the stated conditions: despite optimizing a joint supervised–distillation objective, FORTE achieves a strictly stronger contraction of the supervised task loss than supervised gradient descent at the same admissible

step size. Thus, the distinction lies not only in the practical problem—heterogeneous molecular supervision with both scale imbalance and directional conflict, but also in the theoretical objective: FORTE shows that distillation can accelerate optimization of the downstream task that ultimately determines the deployed model.

A.4 POSITIONING FORTE

Prior representation methods explain why heterogeneous teachers can provide useful information. Multi-teacher methods explain how that information can be combined. Gradient-aware auxiliary methods explain how an auxiliary direction can be compared with a primary task. What remains is a unified rule for heterogeneous distillation that makes teacher participation conditional on task compatibility and evaluates the benefit on the primary objective itself.

FORTE fills this gap by treating heterogeneous distillation as a routing problem. It admits teacher directions according to their first-order compatibility with the downstream gradient, calibrates their influence, and allows the auxiliary route to abstain. The resulting formulation preserves 2D-only inference while making the relevant theoretical target explicit: useful auxiliary supervision should accelerate convergence of the supervised molecular property loss, rather than merely reduce a combined training loss.

B Full-Data and Mini-Batch Parameter-Space Analysis

This section replaces a representation-as-optimization-variable argument by an analysis of the actual empirical objective and the actual GNN parameter update. The representation gradients remain useful only to construct the router and to establish a batch-level certificate. After a single restricted-isometry transfer lemma, every descent and convergence argument is stated directly in parameter space.

B.1 Empirical objective, full-data direction, and actual mini-batch update

Let the finite molecular training set be $\mathcal{D}_N = \{(G_i, y_i)\}_{i=1}^N$. The GNN encoder is $\boldsymbol{h}_i(\theta) = f_\theta(G_i) \in \mathbb{R}^d$, the task head has parameters ϕ, and $w = (\theta, \phi)$ denotes the deployed student parameters. The full empirical supervised objective is

$$F_N(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w), \qquad f_i(w) = \ell_{\sup}(q_\phi(\boldsymbol{h}_i(\theta)), y_i). \tag{15}$$

At iteration t, all gradients below are evaluated at the current parameters; the subscript t is suppressed when no ambiguity arises.

For each molecule, define the task and teacher representation gradients

$$\boldsymbol{g}_i = \nabla_{\boldsymbol{h}_i} f_i, \qquad \boldsymbol{u}_{k,i} = \nabla_{\boldsymbol{h}_i} D_{k,i}. \tag{16}$$

The unchanged FORTE router uses these quantities, the available-only EMA, the cosine mask, the evidence softmax, and the null mass to construct a stopped auxiliary representation gradient $\boldsymbol{a}_i = \sum_{k\in\mathcal{K}} \pi_{k,i}\bar{\boldsymbol{u}}_{k,i}$, as defined in Eq. (10). Its previously established per-molecule certificate is

$$\boldsymbol{g}_i^\top \boldsymbol{a}_i = r_i^2\|\boldsymbol{g}_i\|_2^2, \qquad \|\boldsymbol{a}_i\|_2 \le r_i\|\boldsymbol{g}_i\|_2, \qquad 0 \le r_i \le 1. \tag{17}$$

For any nonempty index set $B \subseteq [N]$ of cardinality b, define the scaled stacked vectors

$$\boldsymbol{g}_B = \frac{1}{\sqrt{b}}(\boldsymbol{g}_i)_{i\in B}, \qquad \boldsymbol{a}_B = \frac{1}{\sqrt{b}}(\boldsymbol{a}_i)_{i\in B}. \tag{18}$$

Let

$$R_B^2 = \frac{\sum_{i\in B} r_i^2\|\boldsymbol{g}_i\|_2^2}{\sum_{i\in B}\|\boldsymbol{g}_i\|_2^2}, \tag{19}$$

with $R_B = 0$ when the denominator vanishes.

Lemma B.1 (Batch lifting of the unchanged router)**.** *For every nonempty $B \subseteq [N]$,*

$$\boldsymbol{g}_B^\top \boldsymbol{a}_B = R_B^2\|\boldsymbol{g}_B\|_2^2, \qquad \|\boldsymbol{a}_B\|_2 \le R_B\|\boldsymbol{g}_B\|_2. \tag{20}$$

Proof. By the definition of the stacked inner product and Equation (17),

$$\boldsymbol{g}_B^\top \boldsymbol{a}_B = \frac{1}{b}\sum_{i\in B} \boldsymbol{g}_i^\top \boldsymbol{a}_i \tag{21}$$

$$= \frac{1}{b}\sum_{i\in B} r_i^2\|\boldsymbol{g}_i\|_2^2 \tag{22}$$

$$= R_B^2 \frac{1}{b}\sum_{i\in B}\|\boldsymbol{g}_i\|_2^2 = R_B^2\|\boldsymbol{g}_B\|_2^2. \tag{23}$$

For the norm,

$$\|\boldsymbol{a}_B\|_2^2 = \frac{1}{b}\sum_{i\in B}\|\boldsymbol{a}_i\|_2^2 \tag{24}$$

$$\le \frac{1}{b}\sum_{i\in B} r_i^2\|\boldsymbol{g}_i\|_2^2 = R_B^2\|\boldsymbol{g}_B\|_2^2. \tag{25}$$

Taking square roots completes the proof. □

Let $\boldsymbol{J}_i = D_\theta \boldsymbol{h}_i(\theta)$ and define the scaled batch Jacobian

$$\boldsymbol{J}_B = \frac{1}{\sqrt{b}} [\boldsymbol{J}_i]_{i \in B}. \tag{26}$$

The task and routed auxiliary gradients on the shared GNN parameters are

$$G_{\theta,B} = \boldsymbol{J}_B^\top \boldsymbol{g}_B = \frac{1}{b} \sum_{i \in B} \boldsymbol{J}_i^\top \boldsymbol{g}_i, \qquad A_{\theta,B} = \boldsymbol{J}_B^\top \boldsymbol{a}_B = \frac{1}{b} \sum_{i \in B} \boldsymbol{J}_i^\top \boldsymbol{a}_i. \tag{27}$$

The task-head gradient is $G_{\phi,B} = b^{-1} \sum_{i \in B} \nabla_\phi f_i$. Hence the task and auxiliary gradients on the deployed parameters $w = (\theta, \phi)$ are

$$G_B = (G_{\theta,B}, G_{\phi,B}), \qquad A_B = (A_{\theta,B}, 0). \tag{28}$$

For $B = [N]$, these are the full-data gradients. For a sampled mini-batch, they are the corresponding batch quantities.

B.2 The correct batch-level restricted-isometry event

The relevant isometry is not a pointwise statement about a freely updated representation. It is a property of the stacked Jacobian of the complete set of graphs whose loss is being differentiated.

Definition B.2 (Batch restricted scaled isometry)**.** For a nonempty $B \subseteq [N]$, let

$$\mathcal{S}_B = \operatorname{span}\{\boldsymbol{g}_B, \boldsymbol{a}_B\}. \tag{29}$$

The batch Jacobian satisfies an (α_B, δ_B) restricted scaled isometry on the FORTE routing plane if $\alpha_B > 0$, $0 \le \delta_B < 1$, and

$$\alpha_B(1-\delta_B)\|\boldsymbol{v}\|_2^2 \le \|\boldsymbol{J}_B^\top \boldsymbol{v}\|_2^2 \le \alpha_B(1+\delta_B)\|\boldsymbol{v}\|_2^2 \qquad \text{for every } \boldsymbol{v} \in \mathcal{S}_B. \tag{30}$$

Equivalently, if $\boldsymbol{U}_B$ has orthonormal columns spanning $\mathcal{S}_B$, the at-most-2 $\times$ 2 compressed Gram matrix

$$\boldsymbol{H}_B = \boldsymbol{U}_B^\top \boldsymbol{J}_B \boldsymbol{J}_B^\top \boldsymbol{U}_B \tag{31}$$

has all eigenvalues in $[\alpha_B(1-\delta_B), \alpha_B(1+\delta_B)]$.

Define the piecewise transfer coefficient

$$\chi(R,\delta) = \begin{cases} \dfrac{R(R-\delta)}{1+\delta}, & R \ge \delta, \\ \dfrac{R(R-\delta)}{1-\delta}, & 0 \le R < \delta. \end{cases} \tag{32}$$

The second branch is negative; its denominator is different because multiplying an inequality by a negative number reverses its direction.

Lemma B.3 (Exact transfer from a batch certificate to GNN parameters)**.** *Suppose Definition B.2 holds for a nonempty set B. Then*

$$G_{\theta,B}^\top A_{\theta,B} \ge \chi(R_B, \delta_B)\|G_{\theta,B}\|_2^2, \tag{33}$$

$$\|A_{\theta,B}\|_2 \le R_B \sqrt{\frac{1+\delta_B}{1-\delta_B}} \|G_{\theta,B}\|_2. \tag{34}$$

In particular, if $R_B \ge \delta_B$, then the routed auxiliary gradient is nonconflicting in the actual shared GNN parameters: $G_{\theta,B}^\top A_{\theta,B} \ge 0$.

Proof. On $\mathcal{S}_B$, Definition B.2 is equivalent to

$$\boldsymbol{J}_B \boldsymbol{J}_B^\top = \alpha_B(I + \boldsymbol{E}_B), \qquad \|\boldsymbol{E}_B\|_{\mathrm{op}} \le \delta_B. \tag{35}$$

Using Equations (27) and (35),

$$G_{\theta,B}^{\top} A_{\theta,B} = \boldsymbol{g}_B^{\top} \boldsymbol{J}_B \boldsymbol{J}_B^{\top} \boldsymbol{a}_B \tag{36}$$
$$= \alpha_B \boldsymbol{g}_B^{\top} \boldsymbol{a}_B + \alpha_B \boldsymbol{g}_B^{\top} \boldsymbol{E}_B \boldsymbol{a}_B \tag{37}$$
$$\geq \alpha_B \boldsymbol{g}_B^{\top} \boldsymbol{a}_B - \alpha_B \delta_B \|\boldsymbol{g}_B\|_2 \|\boldsymbol{a}_B\|_2 \tag{38}$$
$$\geq \alpha_B R_B (R_B - \delta_B) \|\boldsymbol{g}_B\|_2^2, \tag{39}$$

where the last line uses Lemma B.1.

Let $C_B = R_B(R_B - \delta_B)$. When $C_B \geq 0$, Definition B.2 gives

$$\alpha_B \|\boldsymbol{g}_B\|_2^2 \geq \frac{\|G_{\theta,B}\|_2^2}{1 + \delta_B}. \tag{40}$$

Multiplying by the nonnegative number C_B and substituting into Equation (39) yields the first branch of Equation (32). When $C_B < 0$, Definition B.2 instead gives

$$\alpha_B \|\boldsymbol{g}_B\|_2^2 \leq \frac{\|G_{\theta,B}\|_2^2}{1 - \delta_B}. \tag{41}$$

Multiplication by $C_B < 0$ reverses the inequality, which gives the second branch. This proves Equation (33).

For the norm,

$$\|A_{\theta,B}\|_2^2 = \|\boldsymbol{J}_B^{\top} \boldsymbol{a}_B\|_2^2 \tag{42}$$
$$\leq \alpha_B (1 + \delta_B) \|\boldsymbol{a}_B\|_2^2 \tag{43}$$
$$\leq \alpha_B (1 + \delta_B) R_B^2 \|\boldsymbol{g}_B\|_2^2, \tag{44}$$

and

$$\|G_{\theta,B}\|_2^2 \geq \alpha_B (1 - \delta_B) \|\boldsymbol{g}_B\|_2^2. \tag{45}$$

Division and square roots prove Equation (34). □

For later use, define

$$\zeta_B = \begin{cases} \|G_{\theta,B}\|_2^2 / \|G_B\|_2^2, & G_B \neq 0, \\ 0, & G_B = 0, \end{cases} \tag{46}$$

and the parameter-space constants

$$c_B = 1 + \lambda \zeta_B \chi(R_B, \delta_B), \tag{47}$$

$$D_B = 1 + \lambda R_B \sqrt{\frac{1 + \delta_B}{1 - \delta_B}}. \tag{48}$$

Corollary B.4 (Geometry of the complete deployed-parameter direction)**.** *Let $d_B = G_B + \lambda A_B$. Under the conditions of Lemma B.3,*

$$G_B^{\top} d_B \geq c_B \|G_B\|_2^2, \qquad \|d_B\|_2 \leq D_B \|G_B\|_2. \tag{49}$$

Proof. Because the auxiliary objective does not depend on the task-head parameters,

$$G_B^{\top} d_B = \|G_{\phi,B}\|_2^2 + G_{\theta,B}^{\top} (G_{\theta,B} + \lambda A_{\theta,B}) \tag{50}$$
$$\geq \|G_{\phi,B}\|_2^2 + [1 + \lambda \chi(R_B, \delta_B)] \|G_{\theta,B}\|_2^2 \tag{51}$$
$$= c_B \|G_B\|_2^2. \tag{52}$$

Moreover,

$$\|d_B\|_2^2 = \|G_{\phi,B}\|_2^2 + \|G_{\theta,B} + \lambda A_{\theta,B}\|_2^2 \tag{53}$$
$$\leq \|G_{\phi,B}\|_2^2 + D_B^2 \|G_{\theta,B}\|_2^2 \leq D_B^2 \|G_B\|_2^2, \tag{54}$$

because $D_B \geq 1$. □

Theorem B.5 (One-step descent of the realized mini-batch objective)**.** *For a realized nonempty mini-batch B, define*

$$F_B(w) = \frac{1}{|B|} \sum_{i \in B} f_i(w). \tag{55}$$

Suppose F_B is β_B-smooth and the realized batch Jacobian satisfies Definition B.2. Let $d_B = G_B + \lambda A_B$ be the direction actually formed from this batch, including the unchanged current-batch EMA update and stopped routing coefficients. If $c_B > 0$ and

$$0 < \eta \le \frac{c_B}{\beta_B D_B^2}, \tag{56}$$

then

$$F_B(w - \eta d_B) \le F_B(w) - \frac{\eta c_B}{2} \|\nabla F_B(w)\|_2^2. \tag{57}$$

Thus a high-probability isometry event for the realized mini-batch gives a high-probability one-step certificate for the objective that generated the actual stochastic direction. Since B_t changes with t, these inequalities cannot be telescoped across iterations; convergence of the fixed full empirical objective is handled separately in Theorem B.41.

Proof. For the realized batch, $G_B = \nabla F_B(w)$. By Corollary B.4,

$$G_B^\top d_B \ge c_B \|G_B\|_2^2, \qquad \|d_B\|_2 \le D_B \|G_B\|_2. \tag{58}$$

Applying β_B-smoothness at $w - \eta d_B$ gives

$$F_B(w - \eta d_B) \le F_B(w) - \eta G_B^\top d_B + \frac{\beta_B \eta^2}{2} \|d_B\|_2^2 \tag{59}$$

$$\le F_B(w) - \eta c_B \|G_B\|_2^2 + \frac{\beta_B \eta^2 D_B^2}{2} \|G_B\|_2^2. \tag{60}$$

Equation (56) makes the last term at most $\eta c_B \|G_B\|_2^2/2$, proving Equation (57). □

B.3 Full-data convergence after the one-time isometry transfer

All quantities in this subsection are full-data quantities, so $B = [N]$. There is no stochastic mini-batch approximation in the update below.

Assumption B.6 (Smooth full empirical task objective)**.** The empirical objective F_N in Equation (15) is β-smooth and bounded below by $F_\star$.

Theorem B.7 (Full-data convergence on a batch-isometry event)**.** *Suppose Assumption B.6 holds. On an event $\mathcal{E}_{\text{iso}}$, suppose that for every $0 \le t < T$ the full-data batch Jacobian satisfies Definition B.2, and define*

$$c_\star = \min_{0 \le t < T} c_{[N],t} > 0, \qquad D_\star = \max_{0 \le t < T} D_{[N],t} < \infty. \tag{61}$$

Use the unchanged full-data routed direction $d_t = G_t + \lambda A_t$ and the update

$$w_{t+1} = w_t - \eta d_t, \qquad 0 < \eta \le \frac{c_\star}{\beta D_\star^2}. \tag{62}$$

Then, on $\mathcal{E}_{\text{iso}}$,

$$F_N(w_{t+1}) \le F_N(w_t) - \frac{\eta c_\star}{2} \|\nabla F_N(w_t)\|_2^2, \tag{63}$$

and

$$\min_{0 \le t < T} \|\nabla F_N(w_t)\|_2^2 \le \frac{2\beta D_\star^2 [F_N(w_0) - F_\star]}{c_\star^2 T}. \tag{64}$$

Consequently, if $\mathbb{P}(\mathcal{E}_{\text{iso}}) \ge 1 - \delta_{\text{iso}}$, Equation (64) is a $(1 - \delta_{\text{iso}})$-probability full-data convergence guarantee.

Proof. On $\mathcal{E}_{\text{iso}}$, Corollary B.4 gives

$$G_t^\top d_t \geq c_\star \|G_t\|_2^2, \qquad \|d_t\|_2 \leq D_\star \|G_t\|_2, \qquad G_t = \nabla F_N(w_t). \tag{65}$$

By β-smoothness,

$$F_N(w_{t+1}) \leq F_N(w_t) - \eta G_t^\top d_t + \frac{\beta\eta^2}{2}\|d_t\|_2^2 \tag{66}$$

$$\leq F_N(w_t) - \eta c_\star \|G_t\|_2^2 + \frac{\beta\eta^2 D_\star^2}{2}\|G_t\|_2^2. \tag{67}$$

The step-size condition implies $\beta\eta D_\star^2/2 \leq c_\star/2$, which proves Equation (63). Rearranging and summing gives

$$\frac{\eta c_\star}{2}\sum_{t=0}^{T-1}\|G_t\|_2^2 \leq \sum_{t=0}^{T-1}[F_N(w_t) - F_N(w_{t+1})] \tag{68}$$

$$= F_N(w_0) - F_N(w_T) \tag{69}$$

$$\leq F_N(w_0) - F_\star. \tag{70}$$

Divide by $\eta c_\star T/2$, use $\min_t \|G_t\|_2^2 \leq T^{-1}\sum_t \|G_t\|_2^2$, and substitute the largest allowed step size to obtain Equation (64). □

B.4 Full-data rate comparison with uniform and learned teacher weighting

This subsection compares the unchanged FORTE update with uniform aggregation and with two learned-weight settings. Every vector below is an actual parameter gradient on the deployed student parameters $w = (\theta, \phi)$, after the full-batch Jacobian transfer in Lemma B.3. No representation is treated as an independent optimization variable.

Let

$$G_t = \nabla F_N(w_t), \qquad d_t^{\mathsf{M}} = G_t + \lambda V_t^{\mathsf{M}}, \qquad w_{t+1} = w_t - \eta d_t^{\mathsf{M}}, \tag{71}$$

where V_t^{M} is the full-data auxiliary parameter field generated by method M. Two geometrically different regimes are needed.

Definition B.8 (Relative and additive auxiliary regimes)**.** A method is in the *relative regime* on a trajectory if there exist $\gamma_{\mathsf{M}} \geq 0$ and $\kappa_{\mathsf{M}} \geq 0$ such that, for every $t < T$,

$$G_t^\top V_t^{\mathsf{M}} \geq -\gamma_{\mathsf{M}}\|G_t\|_2^2, \qquad \|V_t^{\mathsf{M}}\|_2 \leq \kappa_{\mathsf{M}}\|G_t\|_2. \tag{72}$$

A method is in the *additive regime* if one only knows

$$\|V_t^{\mathsf{M}}\|_2 \leq B_{\mathsf{M}}, \tag{73}$$

where B_{M} need not vanish when $G_t = 0$.

The relative regime preserves the stationary set of the task objective. The additive regime does not: even at $G_t = 0$, a nonzero teacher field may continue to move the student.

Theorem B.9 (Full-data rate in the relative regime)**.** *Suppose Assumption B.6 and Equation (72) hold. Define*

$$c_{\mathsf{M}} = 1 - \lambda\gamma_{\mathsf{M}}, \qquad D_{\mathsf{M}} = 1 + \lambda\kappa_{\mathsf{M}}. \tag{74}$$

If $c_{\mathsf{M}} > 0$ and

$$0 < \eta \leq \frac{c_{\mathsf{M}}}{\beta D_{\mathsf{M}}^2}, \tag{75}$$

then

$$\min_{0\leq t<T}\|\nabla F_N(w_t)\|_2^2 \leq \frac{2\beta D_{\mathsf{M}}^2[F_N(w_0) - F_\star]}{c_{\mathsf{M}}^2 T}. \tag{76}$$

Thus every relative method has deterministic nonconvex order $\mathcal{O}(T^{-1})$, but its leading constant diverges as $\lambda\gamma_{\mathsf{M}} \uparrow 1$.

Proof. By Equation (72),

$$G_t^\top d_t^{\mathsf{M}} = G_t^\top (G_t + \lambda V_t^{\mathsf{M}}) \tag{77}$$
$$= \|G_t\|_2^2 + \lambda G_t^\top V_t^{\mathsf{M}} \tag{78}$$
$$\geq (1 - \lambda\gamma_{\mathsf{M}})\|G_t\|_2^2 \tag{79}$$
$$= c_{\mathsf{M}}\|G_t\|_2^2. \tag{80}$$

Also,

$$\|d_t^{\mathsf{M}}\|_2 = \|G_t + \lambda V_t^{\mathsf{M}}\|_2 \tag{81}$$
$$\leq \|G_t\|_2 + \lambda\|V_t^{\mathsf{M}}\|_2 \tag{82}$$
$$\leq (1 + \lambda\kappa_{\mathsf{M}})\|G_t\|_2 \tag{83}$$
$$= D_{\mathsf{M}}\|G_t\|_2. \tag{84}$$

By β-smoothness of the fixed full empirical objective,

$$F_N(w_{t+1}) \leq F_N(w_t) - \eta G_t^\top d_t^{\mathsf{M}} + \frac{\beta\eta^2}{2}\|d_t^{\mathsf{M}}\|_2^2 \tag{85}$$
$$\leq F_N(w_t) - \left(\eta c_{\mathsf{M}} - \frac{\beta\eta^2 D_{\mathsf{M}}^2}{2}\right)\|G_t\|_2^2. \tag{86}$$

The step-size condition gives

$$\eta c_{\mathsf{M}} - \frac{\beta\eta^2 D_{\mathsf{M}}^2}{2} \geq \frac{\eta c_{\mathsf{M}}}{2}, \tag{87}$$

so

$$\frac{\eta c_{\mathsf{M}}}{2}\|G_t\|_2^2 \leq F_N(w_t) - F_N(w_{t+1}). \tag{88}$$

Summing over $t = 0, \ldots, T-1$ telescopes the same objective F_N:

$$\frac{\eta c_{\mathsf{M}}}{2}\sum_{t=0}^{T-1}\|G_t\|_2^2 \leq \sum_{t=0}^{T-1}[F_N(w_t) - F_N(w_{t+1})] \tag{89}$$
$$= F_N(w_0) - F_N(w_T) \tag{90}$$
$$\leq F_N(w_0) - F_\star. \tag{91}$$

Divide by $\eta c_{\mathsf{M}} T/2$, apply $\min_t \|G_t\|_2^2 \leq T^{-1}\sum_t \|G_t\|_2^2$, and substitute the largest admissible step in Equation (75). □

Theorem B.10 (A nonvanishing auxiliary field creates a stationarity floor)**.** *Suppose Assumption B.6 and Equation (73) hold. Use $\eta = 1/(2\beta)$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_N(w_t)\|_2^2 \leq \frac{16\beta[F_N(w_0) - F_\star]}{3T} + \frac{10}{9}\lambda^2 B_{\mathsf{M}}^2. \tag{92}$$

Hence the general order is $\mathcal{O}(T^{-1}) + \mathcal{O}(\lambda^2 B_{\mathsf{M}}^2)$ rather than exact convergence to task stationarity.

Proof. Set $e_t = \lambda V_t^{\mathsf{M}}$, so $\|e_t\|_2 \leq \lambda B_{\mathsf{M}}$ and $d_t^{\mathsf{M}} = G_t + e_t$. Smoothness gives

$$F_N(w_{t+1}) \leq F_N(w_t) - \eta G_t^\top (G_t + e_t) + \frac{\beta\eta^2}{2}\|G_t + e_t\|_2^2 \tag{93}$$
$$= F_N(w_t) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|G_t\|_2^2 - \eta(1 - \beta\eta)G_t^\top e_t + \frac{\beta\eta^2}{2}\|e_t\|_2^2. \tag{94}$$

Let

$$a = 1 - \frac{\beta\eta}{2}, \qquad q = 1 - \beta\eta. \tag{95}$$

For $0 < \eta \le 1/\beta$, $a > 0$ and $q \ge 0$. By Cauchy–Schwarz and Young's inequality,

$$q|G_t^\top e_t| \le q\|G_t\|_2\|e_t\|_2 \tag{96}$$

$$\le \frac{a}{2}\|G_t\|_2^2 + \frac{q^2}{2a}\|e_t\|_2^2. \tag{97}$$

Substitution into Equation (94) yields

$$F_N(w_{t+1}) \le F_N(w_t) - \frac{\eta a}{2}\|G_t\|_2^2 + \frac{\eta}{2}\left(\frac{q^2}{a} + \beta\eta\right)\|e_t\|_2^2. \tag{98}$$

For $\eta = 1/(2\beta)$, $a = 3/4$ and $q = 1/2$. Sum Equation (98), telescope F_N, use $\|e_t\|_2^2 \le \lambda^2 B_{\mathsf{M}}^2$, and divide by $\eta a T/2$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|G_t\|_2^2 \le \frac{2[F_N(w_0) - F_\star]}{\eta a T} + \left(\frac{q^2}{a^2} + \frac{\beta\eta}{a}\right)\lambda^2 B_{\mathsf{M}}^2 \tag{99}$$

$$= \frac{16\beta[F_N(w_0) - F_\star]}{3T} + \frac{10}{9}\lambda^2 B_{\mathsf{M}}^2. \tag{100}$$

□

Definition B.11 (The four comparison methods)**.** The following baselines use the same teachers and deployed GNN as FORTE. Let $U_{k,t}$ denote teacher k's full-data auxiliary parameter gradient.

(i) *Raw uniform:*

$$V_t^{\text{raw-uni}} = \frac{1}{|\mathcal{A}_t|}\sum_{k\in\mathcal{A}_t} U_{k,t}, \tag{101}$$

with no task-relative gradient calibration.

(ii) *Calibrated uniform:* the same task-relative calibrated teacher gradients as FORTE are averaged uniformly, but no cosine mask, evidence weighting, or null gate is used.

(iii) *Detached learned weighting:* a learned global, batch-dependent, or graph-dependent simplex vector $\alpha_{k,i}(\omega_t)$ is treated as a constant in the student backward pass. Its student auxiliary field is the corresponding convex combination of teacher parameter gradients.

(iv) *End-to-end learned gating:* the gate is differentiated through. For the full-data weighted teacher objective, the product rule gives

$$V_t^{\text{e2e}} = \underbrace{\frac{1}{N}\sum_{i,k}\alpha_{k,i}U_{k,i}}_{V_t^{\text{value}}} + \underbrace{\frac{1}{N}\sum_{i,k}D_{k,i}\nabla_w\alpha_{k,i}}_{H_t^{\text{gate}}}. \tag{102}$$

The two learned settings differ because end-to-end differentiation introduces the additional gate-Jacobian field H_t^{gate}.

Corollary B.12 (Deterministic order comparison)**.** *On the full-data batch-isometry event, FORTE has*

$$c_{\text{F}} = \min_{t<T} c_{[N],t} > 0, \qquad D_{\text{F}} = \max_{t<T} D_{[N],t}, \tag{103}$$

and therefore

$$\min_{t<T}\|G_t\|_2^2 \le \frac{2\beta D_{\text{F}}^2\Delta_0}{c_{\text{F}}^2 T}, \qquad \Delta_0 = F_N(w_0) - F_\star. \tag{104}$$

For calibrated uniform or calibrated detached learned weighting, let $\overline{\delta}_{\mathsf{M}} < 1$ bound the batch restricted-isometry distortion on that comparator's own plane $\operatorname{span}\{\boldsymbol{g}_{[N]}, \boldsymbol{v}_{\mathsf{M},[N]}\}$. *Then one may*

take

$$\kappa_{\mathsf{M}} = \sqrt{\frac{1+\bar{\delta}_{\mathsf{M}}}{1-\bar{\delta}_{\mathsf{M}}}}. \tag{105}$$

If, in addition,

$$G_t^\top V_t^{\mathsf{M}} \geq -\gamma_{\mathsf{M}} \|G_t\|_2^2, \qquad \lambda\gamma_{\mathsf{M}} < 1, \tag{106}$$

then

$$\min_{t<T} \|G_t\|_2^2 \leq \frac{2\beta(1+\lambda\kappa_{\mathsf{M}})^2\Delta_0}{(1-\lambda\gamma_{\mathsf{M}})^2 T}. \tag{107}$$

The ratio of the two upper-bound coefficients *is exactly*

$$\frac{C_{\mathsf{M}}}{C_{\mathrm{F}}} = \frac{(1+\lambda\kappa_{\mathsf{M}})^2 c_{\mathrm{F}}^2}{(1-\lambda\gamma_{\mathsf{M}})^2 D_{\mathrm{F}}^2}. \tag{108}$$

Raw uniform and an uncalibrated detached learned gate generally satisfy only Theorem B.10. End-to-end learned gating also falls in the additive regime whenever the complete field $V_t^{\mathrm{value}} + H_t^{\mathrm{gate}}$ is bounded but does not vanish with $\|G_t\|_2$. It enters the relative regime only if the entire *vector in Equation (102) satisfies Equation (72).*

Proof. Equation (104) is Theorem B.7. For a calibrated convex combination, let $v_{\mathsf{M},i} = \sum_k \alpha_{k,i} \bar{\boldsymbol{u}}_{k,i}$, where $\alpha_{k,i} \geq 0$, $\sum_k \alpha_{k,i} = 1$, and $\|\bar{\boldsymbol{u}}_{k,i}\|_2 \leq \|\boldsymbol{g}_i\|_2$. Then

$$\|v_{\mathsf{M},i}\|_2 \leq \sum_k \alpha_{k,i} \|\bar{\boldsymbol{u}}_{k,i}\|_2 \tag{109}$$

$$\leq \|\boldsymbol{g}_i\|_2. \tag{110}$$

After stacking over the full dataset,

$$\|\boldsymbol{v}_{\mathsf{M},[N]}\|_2 \leq \|\boldsymbol{g}_{[N]}\|_2. \tag{111}$$

Apply the comparator-specific batch restricted isometry to $V_{\theta,t}^{\mathsf{M}} = \mathcal{J}_{[N]}^\top \boldsymbol{v}_{\mathsf{M},[N]}$ and $G_{\theta,t} = \mathcal{J}_{[N]}^\top \boldsymbol{g}_{[N]}$:

$$\|V_{\theta,t}^{\mathsf{M}}\|_2^2 \leq \alpha_t(1+\bar{\delta}_{\mathsf{M}}) \|\boldsymbol{v}_{\mathsf{M},[N]}\|_2^2 \tag{112}$$

$$\leq \alpha_t(1+\bar{\delta}_{\mathsf{M}}) \|\boldsymbol{g}_{[N]}\|_2^2 \tag{113}$$

$$\leq \frac{1+\bar{\delta}_{\mathsf{M}}}{1-\bar{\delta}_{\mathsf{M}}} \|G_{\theta,t}\|_2^2 \tag{114}$$

$$\leq \kappa_{\mathsf{M}}^2 \|G_t\|_2^2. \tag{115}$$

This proves Equation (105). Under Equation (106), Theorem B.9 gives Equation (107). Dividing the coefficients multiplying $1/T$ gives Equation (108).

Without task-relative calibration, a bounded auxiliary field need not vanish at $G_t = 0$, so Theorem B.10 applies. For an end-to-end gate, the product rule gives Equation (102); controlling only the convex-combination term does not control the actual student update because H_t^{gate} can have an arbitrary task projection. □

Corollary B.13 (Order table)**.** *The worst-case full-data nonconvex orders implied above are*

Method and condition	*Task-gradient bound*
FORTE on the batch-isometry event	$\mathcal{O}(T^{-1})$
Calibrated uniform with $\lambda\gamma_{\mathrm{U}} < 1$	$\mathcal{O}(T^{-1})$
Calibrated detached learned gate with $\lambda\gamma_{\mathrm{L}} < 1$	$\mathcal{O}(T^{-1})$
Raw uniform or an uncalibrated detached gate	$\mathcal{O}(T^{-1}) + \mathcal{O}(\lambda^2 B^2)$
End-to-end learned gate without whole-field relative geometry	$\mathcal{O}(T^{-1}) + \mathcal{O}(\lambda^2 B^2)$

Thus FORTE's order-level advantage over raw uniform and unconstrained learned gating is removal of the nonvanishing T^0 stationarity floor. Relative calibrated comparators retain the same

T^{-1} exponent, but only under an external conflict-margin condition and with a conflict-dependent constant.

Proof. Apply Theorem B.9 to the relative rows and Theorem B.10 to the additive rows. □

Corollary B.14 (PL comparison)**.** *Suppose additionally that*

$$\frac{1}{2}\|\nabla F_N(w)\|_2^2 \geq \mu[F_N(w) - F_\star]. \tag{116}$$

A relative method with the largest step in Equation (75) satisfies

$$F_N(w_T) - F_\star \leq \left(1 - \frac{\mu c_{\mathsf{M}}^2}{\beta D_{\mathsf{M}}^2}\right)^T [F_N(w_0) - F_\star]. \tag{117}$$

An additive method with $\eta = 1/(2\beta)$ satisfies

$$F_N(w_T) - F_\star \leq \left(1 - \frac{3\mu}{8\beta}\right)^T \Delta_0 + \frac{5\lambda^2 B_{\mathsf{M}}^2}{9\mu}. \tag{118}$$

Thus relative methods converge linearly to the task optimum, whereas a nonvanishing raw uniform or learned auxiliary field is guaranteed only to converge linearly to a fixed neighborhood.

Proof. The proof of Theorem B.9 gives

$$F_N(w_{t+1}) - F_\star \leq F_N(w_t) - F_\star - \frac{\eta c_{\mathsf{M}}}{2}\|G_t\|_2^2. \tag{119}$$

Using Equation (116) and $\eta = c_{\mathsf{M}}/(\beta D_{\mathsf{M}}^2)$ gives Equation (117).

For an additive method, Equation (98) at $\eta = 1/(2\beta)$ becomes

$$F_N(w_{t+1}) - F_\star \leq F_N(w_t) - F_\star - \frac{3\eta}{8}\|G_t\|_2^2 + \frac{5\eta}{12}\lambda^2 B_{\mathsf{M}}^2. \tag{120}$$

The PL inequality gives

$$\Delta_{t+1} \leq \left(1 - \frac{3\mu}{8\beta}\right)\Delta_t + \frac{5}{24\beta}\lambda^2 B_{\mathsf{M}}^2. \tag{121}$$

Unroll the recursion and use $\sum_{j=0}^{T-1}(1-q)^j \leq q^{-1}$ with $q = 3\mu/(8\beta)$ to obtain Equation (118). □

B.4.1 Distillation-accelerated convergence of the supervised task loss

The preceding comparison controls task-gradient stationarity. We now retain the task–auxiliary cross term to obtain a stronger contraction guarantee for the supervised task loss itself. Throughout this subsection, F_N, $w = (\boldsymbol{\theta}, \boldsymbol{\phi})$, $G_t = \nabla F_N(w_t)$, and $A_t = (A_{\boldsymbol{\theta},t}, 0)$ are exactly the quantities defined in Section B.1. The update remains $w_{t+1} = w_t - \eta(G_t + \lambda A_t)$, and $\Delta_t = F_N(w_t) - F_\star$. Here $F_\star$ is the task optimum for the PL analysis. We do not require the iteration-dependent distillation objective to decrease or to share an optimum with F_N.

Correspondence with the existing geometry. Write $R_t = R_{[N],t}$, $\delta_t = \delta_{[N],t}$, $\alpha_t = \alpha_{[N],t}$, and $\xi_t = \xi_{[N],t}$, where

$$R_t^2 = \frac{\sum_i r_{i,t}^2 \|\boldsymbol{g}_{i,t}\|_2^2}{\sum_i \|\boldsymbol{g}_{i,t}\|_2^2}, \qquad \xi_t = \frac{\|G_{\boldsymbol{\theta},t}\|_2^2}{\|G_t\|_2^2}. \tag{122}$$

The zero-denominator conventions are those in Equations (19) and (46). In particular, ξ_t is the encoder's share of squared task-gradient norm, not the per-molecule task-gradient norm ζ_i in Equation (7). All stacked vectors and Jacobians retain the $N^{-1/2}$ scaling.

Lemma B.15 (Auxiliary norm controlled by task-alignment gain)**.** *Suppose Definition B.2 holds on the full-data routing plane and $R_t > \delta_t$. Define*

$$m_t = \xi_t \frac{R_t(R_t - \delta_t)}{1 + \delta_t}, \qquad \Lambda_t = \frac{(1+\delta_t)R_t}{R_t - \delta_t}. \tag{123}$$

Then

$$\langle G_t, A_t\rangle \geq m_t\|G_t\|_2^2, \qquad \|A_t\|_2^2 \leq \Lambda_t\langle G_t, A_t\rangle. \tag{124}$$

Here m_t and Λ_t are derived geometric constants, not additional routing hyperparameters; m_t is distinct from the availability mask $m_{k,i}$. Equation (47) gives $c_{[N],t} = 1 + \lambda m_t$ in this regime.

Proof. Let $\boldsymbol{g}_t = \boldsymbol{g}_{[N],t}$, $\boldsymbol{a}_t = \boldsymbol{a}_{[N],t}$, and $\boldsymbol{J}_t = \boldsymbol{J}_{[N],t}$. Lemma B.1 gives $\boldsymbol{g}_t^\top \boldsymbol{a}_t = R_t^2\|\boldsymbol{g}_t\|_2^2$ and $\|\boldsymbol{a}_t\|_2^2 \leq R_t^2\|\boldsymbol{g}_t\|_2^2$. Applying the compressed-operator identity in Equation (35) on the routing plane yields

$$\begin{aligned}
\langle G_t, A_t\rangle &= \boldsymbol{g}_t^\top \boldsymbol{J}_t\boldsymbol{J}_t^\top \boldsymbol{a}_t && (125)\\
&\geq \alpha_t\big(\boldsymbol{g}_t^\top \boldsymbol{a}_t - \delta_t\|\boldsymbol{g}_t\|_2\|\boldsymbol{a}_t\|_2\big) && (126)\\
&\geq \alpha_t R_t(R_t - \delta_t)\|\boldsymbol{g}_t\|_2^2, && (127)\\
\|A_t\|_2^2 = \|\boldsymbol{J}_t^\top \boldsymbol{a}_t\|_2^2 &\leq \alpha_t(1+\delta_t)R_t^2\|\boldsymbol{g}_t\|_2^2. && (128)
\end{aligned}$$

Combining these inequalities proves the norm–gain coupling without dividing by an inner product that might vanish. The isometry upper bound $\|G_{\boldsymbol{\theta},t}\|_2^2 \leq \alpha_t(1+\delta_t)\|\boldsymbol{g}_t\|_2^2$ and the definition of ξ_t give the first claim. If $G_t = 0$, restricted isometry implies $\boldsymbol{g}_t = 0$ and hence $A_t = 0$, so the inequalities remain valid by the zero conventions. When $\delta_t = 0$ and $R_t > 0$, the coupling reduces to $\Lambda_t = 1$ and $m_t = \xi_t R_t^2$. □

Theorem B.16 (Stronger task-loss contraction at a common step size)**.** *Assume F_N is β-smooth and satisfies the PL condition in Equation (116), with constant $\mu > 0$, on a domain containing the analyzed FORTE and supervised trajectories and the update segments. Suppose Equation (124) holds along the FORTE trajectory with uniform bounds $m_t \geq m > 0$ and $\Lambda_t \leq \Lambda < \infty$ at nonstationary iterates. For $\lambda > 0$, choose a common step size*

$$0 < \eta < \frac{1}{\beta(1+\lambda\Lambda)}, \qquad h_0 = 1 - \frac{\beta\eta}{2}, \qquad \vartheta_\eta = 1 - \beta\eta\left(1 + \frac{\lambda\Lambda}{2}\right) > 0, \tag{129}$$

and set $H_{\mathrm{F}} = h_0 + \lambda\vartheta_\eta m$. Also require $0 < 2\mu\eta H_{\mathrm{F}} < 1$. Then

$$\Delta_T \leq (1 - 2\mu\eta H_{\mathrm{F}})^T\Delta_0. \tag{130}$$

Supervised gradient descent from the same initial parameters with the same η satisfies

$$F_N(w_T^{\mathrm{sup}}) - F_\star \leq (1 - 2\mu\eta h_0)^T\Delta_0. \tag{131}$$

Since $H_{\mathrm{F}} > h_0$, FORTE has a strictly smaller contraction factor in this guarantee.

Proof. Smoothness of the fixed supervised objective gives

$$\begin{aligned}
F_N(w_{t+1}) &\leq F_N(w_t) - \eta\langle G_t, G_t + \lambda A_t\rangle + \frac{\beta\eta^2}{2}\|G_t + \lambda A_t\|_2^2 && (132)\\
&= F_N(w_t) - \eta h_0\|G_t\|_2^2 - \eta\lambda(1-\beta\eta)\langle G_t, A_t\rangle + \frac{\beta\eta^2\lambda^2}{2}\|A_t\|_2^2 && (133)\\
&\leq F_N(w_t) - \eta h_0\|G_t\|_2^2 - \eta\lambda\vartheta_\eta\langle G_t, A_t\rangle && (134)\\
&\leq F_N(w_t) - \eta H_{\mathrm{F}}\|G_t\|_2^2. && (135)
\end{aligned}$$

The penultimate line uses $\|A_t\|_2^2 \leq \Lambda\langle G_t, A_t\rangle$; the final line uses $\vartheta_\eta > 0$ and $\langle G_t, A_t\rangle \geq m\|G_t\|_2^2$. Thus the retained cross term provides an extra decrease of the supervised loss, not merely of the distillation loss. Applying $\|G_t\|_2^2 \geq 2\mu\Delta_t$ and iterating proves Equation (130). Setting $A_t = 0$ gives the supervised bound. These are comparisons of guarantees; they do not order the actual losses of two different trajectories at every iteration. □

Sufficient iteration budgets. For $0 < \epsilon_{\mathrm{opt}} < \Delta_0$, the inequality $1 - x \leq e^{-x}$ gives the sufficient budgets

$$\widehat{T}_{\mathrm{F}}(\epsilon_{\mathrm{opt}}) = \left\lceil \frac{\log(\Delta_0/\epsilon_{\mathrm{opt}})}{2\mu\eta H_{\mathrm{F}}} \right\rceil, \qquad \widehat{T}_{\mathrm{sup}}(\epsilon_{\mathrm{opt}}) = \left\lceil \frac{\log(\Delta_0/\epsilon_{\mathrm{opt}})}{2\mu\eta h_0} \right\rceil. \tag{136}$$

Ignoring integer rounding, the improvement factor is

$$S_{\text{task}} = \frac{H_{\text{F}}}{h_0} = 1 + \frac{\lambda \vartheta_\eta m}{1 - \beta\eta/2} > 1. \tag{137}$$

Integer budgets can coincide after rounding. The target accuracy ϵ_{opt} is distinct from the numerical stabilizers $\varepsilon_c, \varepsilon_\omega$ and the stochastic direction radius ε_t.

Corollary B.17 (A routing-based sufficient condition for strict gain)**.** *Suppose, on the nonstationary trajectory, $R_t \geq R_{\min} > \overline{\delta} \geq \delta_t$ and $\xi_t \geq \xi_{\min} > 0$. Then the uniform constants in Theorem B.16 can be chosen as*

$$m = \xi_{\min} \frac{R_{\min}(R_{\min} - \overline{\delta})}{1 + \overline{\delta}} > 0, \qquad \Lambda = \frac{(1 + \overline{\delta})R_{\min}}{R_{\min} - \overline{\delta}} < \infty. \tag{138}$$

Proof. For $R > \delta$, $R(R-\delta)/(1+\delta)$ increases with R and decreases with δ, whereas $(1+\delta)R/(R-\delta)$ decreases with R and increases with δ. Substitution into Equation (123) proves both bounds. The condition on ξ_t is necessary for this uniform guarantee: distillation does not update the task head directly. □

Corollary B.18 (Accumulated task gain with null-route iterations)**.** *Allow $m_t \geq 0$ instead of a uniform positive lower bound. Suppose each iteration either has $A_t = 0$ or satisfies Equation (124) with a common finite Λ; set $m_t = 0$ on null iterations. Under the smoothness, PL, and step-size conditions above, let $q_{\text{sup}} = 1 - 2\mu\eta h_0 \in (0, 1)$ and assume $0 < q_{\text{sup}} - 2\mu\eta\lambda\vartheta_\eta m_t < 1$ for every $t < T$. Then*

$$\Delta_T \leq \prod_{t=0}^{T-1} (q_{\text{sup}} - 2\mu\eta\lambda\vartheta_\eta m_t)\Delta_0 \tag{139}$$

$$\leq q_{\text{sup}}^T \exp\left(-\frac{2\mu\eta\lambda\vartheta_\eta}{q_{\text{sup}}} \sum_{t=0}^{T-1} m_t \right) \Delta_0. \tag{140}$$

Proof. The proof of Equation (135) applies with m replaced by m_t; on a null iteration it is the supervised descent bound. Iterate the resulting PL inequality and apply $1 - x \leq e^{-x}$ to each factor after factoring out q_{sup}. Thus null iterations recover the supervised guarantee, and helpful iterations accumulate additional progress. A positive long-run average of m_t gives a sustained improvement in the exponential bound; finitely many helpful iterations alone do not establish an asymptotic rate improvement. □

Proposition B.19 (Actual one-step task-loss comparison from a common state)**.** *At a common nonstationary state w, let $G = \nabla F_N(w)$ and suppose $\langle G, A\rangle \geq m\|G\|_2^2$, $\|A\|_2^2 \leq \Lambda\langle G, A\rangle$ with $m > 0$. For $w_{\text{F}}^+ = w - \eta(G + \lambda A)$ and $w_{\text{sup}}^+ = w - \eta G$, β-smoothness gives*

$$F_N(w_{\text{F}}^+) \leq F_N(w_{\text{sup}}^+) - \eta\lambda \left[1 - \beta\eta \left(\sqrt{\frac{\Lambda}{m}} + \frac{\lambda\Lambda}{2} \right) \right] \langle G, A\rangle. \tag{141}$$

In particular, for $\lambda > 0$ and $0 < \eta < [\beta(\sqrt{\Lambda/m} + \lambda\Lambda/2)]^{-1}$, $F_N(w_{\text{F}}^+) < F_N(w_{\text{sup}}^+)$.

Proof. Expand at w_{sup}^+ and use $\|\nabla F_N(w_{\text{sup}}^+) - G\|_2 \leq \beta\eta\|G\|_2$:

$$F_N(w_{\text{F}}^+) \leq F_N(w_{\text{sup}}^+) - \eta\lambda\langle G, A\rangle + \beta\eta^2\lambda\|G\|_2\|A\|_2 + \frac{\beta\eta^2\lambda^2}{2}\|A\|_2^2. \tag{142}$$

The two geometric assumptions imply $\|G\|_2\|A\|_2 \leq \sqrt{\Lambda/m}\langle G, A\rangle$. Substituting proves the bound, and nonstationarity makes $\langle G, A\rangle > 0$. This one-step comparison requires a common starting state; it does not imply pointwise ordering of separate runs. □

Relation to the existing uniform comparison. Corollary B.14 already gives a task-loss contraction coefficient $\mu c_{\mathsf{M}}^2/(\beta D_{\mathsf{M}}^2)$ at the method-specific step $\eta_{\mathsf{M}} = c_{\mathsf{M}}/(\beta D_{\mathsf{M}}^2)$. Substituting $(c_{\mathrm{F}}, D_{\mathrm{F}})$ for FORTE and $(1-\lambda\gamma_{\mathrm{U}}, 1+\lambda\kappa_{\mathrm{U}})$ for calibrated uniform recovers $s_{\mathrm{F}\leftarrow\mathrm{U}}$ in Equation (14). Under a common PL constant, the same ratio compares the unrounded exponential sufficient budgets for task loss. However, Cauchy–Schwarz gives $c_{\mathrm{F}} \le D_{\mathrm{F}}$ on a nonstationary trajectory. That earlier bound therefore does not establish an advantage over supervised gradient descent with its own optimized step. Theorem B.16 instead retains the cross term and compares both methods at the same explicitly admissible step size.

B.5 GNN REGIMES IN WHICH THE BATCH ISOMETRY HOLDS

The following results concern only graph neural networks. They do not claim that every GNN is automatically isometric. Instead, they identify explicit GNN tangent-feature conditions that imply the event required by Theorem B.7, for either the full dataset or a mini-batch.

B.5.1 CONCRETE GCN, GIN, AND MULTILAYER MESSAGE-PASSING STUDENTS

The batch restricted-isometry condition concerns the complete batch computation. It does not require every graph or every channel to be individually isometric. We now construct three standard graph-level GNNs and show exactly which additional width and graph-kernel conditions yield the batch event with high probability.

Definition B.20 (Three concrete graph-level GNN encoders)**.** For a molecular graph G_i with node-feature matrix X_i and adjacency matrix A_i, consider the following depth-L trunks.

(i) *L-layer GCN.* Let $\widehat{A}_i = \widetilde{D}_i^{-1/2}(A_i + I)\widetilde{D}_i^{-1/2}$ and

$$H_i^{(0)} = X_i, \qquad H_i^{(\ell)} = \sigma_\ell(\widehat{A}_i H_i^{(\ell-1)} W_\ell), \quad 1 \le \ell \le L. \tag{143}$$

This is the multilayer propagation rule of Kipf & Welling (2017, Eq. (2), paper p. 2). A bounded normalized graph pooling operator P_i produces $z_i = P_i H_i^{(L)} \in \mathbb{R}^m$.

(ii) *L-layer GIN.* Let

$$H_i^{(\ell)} = \mathrm{MLP}_\ell\Big([(1+\epsilon_\ell)I + A_i]H_i^{(\ell-1)}\Big), \quad 1 \le \ell \le L, \tag{144}$$

with $H_i^{(0)} = X_i$. This is the GIN update of Xu et al. (2019, Eq. (4.1), paper p. 5). We may use the final-layer graph vector $z_i = P_i H_i^{(L)}$, or the original concatenated layerwise readout of Xu et al. (2019, Eq. (4.2), paper p. 6); the latter is a finite direct sum of the same analysis.

(iii) *Normalized multilayer MPNN.* For node v,

$$m_{i,v}^{(\ell)} = \sum_{u \in \mathcal{N}_i(v)\cup\{v\}} \rho_{i,vu}^{(\ell)} \psi_\ell(H_{i,v}^{(\ell-1)}, H_{i,u}^{(\ell-1)}, e_{vu}), \tag{145}$$

$$H_{i,v}^{(\ell)} = \phi_\ell(H_{i,v}^{(\ell-1)}, m_{i,v}^{(\ell)}), \tag{146}$$

where $\sup_{i,v,\ell} \sum_u |\rho_{i,vu}^{(\ell)}| \le C_A < \infty$. Again set $z_i = P_i H_i^{(L)} \in \mathbb{R}^m$.

In every case the routed representation is produced by the ordinary shared linear graph projection

$$\boldsymbol{h}_i = W_{\mathrm{out}} \frac{z_i}{\sqrt{m}}. \tag{147}$$

The width normalization is used only in the theoretical parameterization; it can equivalently be absorbed into the initialization scale of W_{out}. The FORTE router, EMA, evidence gate, null route, and final objective are unchanged.

Proposition B.21 (Bounded-gradient regime for the three concrete GNNs)**.** *Assume a finite graph-size bound $|V_i| \le n_{\max}$, bounded node and edge features, bounded weight operator norms on the analyzed parameter ball, and Lipschitz activations and MLPs. For GIN also assume $\deg_i(v) \le \Delta_{\max}$ and $\sup_\ell |1+\epsilon_\ell| < \infty$. Suppose the task and teacher heads have bounded Jacobians and bounded loss derivatives. Then each architecture in Definition B.20 has bounded graph embeddings, bounded per-graph encoder Jacobians, bounded routed parameter directions, and bounded teacher gradient-norm observations. Consequently, the conditional Bernstein conditions in Definition B.39 hold by the bounded-increment special case.*

Proof. We verify each network separately.

For GCN, the symmetric normalized propagation operator satisfies $\|\widehat{A}_i\|_{\mathrm{op}} \le 1$. Let $\|W_\ell\|_{\mathrm{op}} \le M_\ell$, let σ_ℓ be L_ℓ-Lipschitz, and define

$$B_0 = \sup_i \|X_i\|_F, \qquad B_\ell = L_\ell M_\ell B_{\ell-1}. \tag{148}$$

Then

$$\|H_i^{(\ell)}\|_F \le L_\ell \|\widehat{A}_i H_i^{(\ell-1)} W_\ell\|_F \tag{149}$$

$$\le L_\ell M_\ell \|H_i^{(\ell-1)}\|_F \le B_\ell. \tag{150}$$

A perturbation inserted at layer s is multiplied by at most $L_\ell M_\ell$ at each later layer. Bounded pooling and finite depth therefore give a finite block-Jacobian bound and, after summing the squared block bounds, a finite $\|D_\theta \boldsymbol{h}_i\|_{\mathrm{op}}$.

For GIN, an undirected graph with maximum degree $\Delta_{\max}$ satisfies $\|A_i\|_{\mathrm{op}} \le \Delta_{\max}$. Hence

$$\|(1+\epsilon_\ell)I + A_i\|_{\mathrm{op}} \le |1+\epsilon_\ell| + \Delta_{\max} =: C_{\mathrm{GIN},\ell}. \tag{151}$$

If MLP_ℓ has Lipschitz constant $L_{\mathrm{MLP},\ell}$, then

$$\|H_i^{(\ell)}\|_F \le L_{\mathrm{MLP},\ell} C_{\mathrm{GIN},\ell} \|H_i^{(\ell-1)}\|_F. \tag{152}$$

A sum readout is bounded by $\sqrt{n_{\max}}\|H_i^{(\ell)}\|_F$. Repeated application of the chain rule through finitely many GIN layers and MLP factors gives a finite encoder Jacobian. Thus an ordinary multilayer GIN is covered; the constants grow with $\Delta_{\max}$ and $n_{\max}$, but no normalization of the GIN sum aggregator is required.

For the normalized MPNN, the aggregation map has Lipschitz constant at most $C_A L_{\psi,\ell}$, and the update map has finite Lipschitz constant $L_{\phi,\ell}$. Thus, for a finite architecture-dependent C_ℓ,

$$\|H_i^{(\ell)}\|_F \le C_\ell(1 + \|H_i^{(\ell-1)}\|_F). \tag{153}$$

Finite depth gives a finite feature bound, and the same perturbation recursion gives a finite Jacobian bound.

In all three cases, bounded task-head and loss derivatives give $\|\boldsymbol{g}_i\|_2 \le B_g$. The unchanged FORTE calibration gives $\|\boldsymbol{a}_i\|_2 \le \|\boldsymbol{g}_i\|_2$. Therefore

$$\|\boldsymbol{J}_i^\top(\boldsymbol{g}_i + \lambda \boldsymbol{a}_i)\|_2 \le \|\boldsymbol{J}_i\|_{\mathrm{op}}(1+\lambda)B_g. \tag{154}$$

The private teacher-head bounds likewise control $\|\boldsymbol{u}_{k,i}\|_2$. The centered finite-population increments are therefore bounded, which implies the Bernstein moment condition used in Proposition B.47. □

Lemma B.22 (Conditional channel independence in standard wide GNNs)**.** *Suppose the final node-feature width is m and the output-channel parameter vectors of the final affine map are independently initialized. Conditional on every preceding layer, the graph-channel vectors*

$$x_{B,r} = (z_{i,r})_{i\in B} \in \mathbb{R}^{|B|}, \qquad 1 \le r \le m, \tag{155}$$

are independent across r for each of the GCN, GIN, and normalized MPNN in Definition B.20.

Proof. For GCN, condition on $H_i^{(L-1)}$ for all graphs. The rth final node channel has the form

$$H_{i,:,r}^{(L)} = \sigma_L(\widehat{A}_i H_i^{(L-1)} w_{L,r}), \tag{156}$$

where $w_{L,r}$ is column r of W_L. The pooled scalar $z_{i,r} = P_i H_{i,:,r}^{(L)}$ is a measurable function only of $w_{L,r}$ and the conditioned quantities. Independent columns $w_{L,r}$ therefore produce independent vectors $x_{B,r}$.

For GIN, expose the final MLP up to its last affine output layer. Conditional on its penultimate activations, output channel r is a measurable function of only column r of the final MLP weight matrix and its corresponding bias. Independent output columns again imply independence of the pooled vectors $x_{B,r}$.

For a normalized MPNN, condition on all messages and on the penultimate hidden features entering the final output affine map of ϕ_L. Each final channel is then a measurable function of its own independently initialized output parameter vector. Pooling preserves channelwise independence. This proves the claim for all three architectures. □

Lemma B.23 (A graph-kernel condition that implies mean batch isometry)**.** *Fix a graph set B of size b. Let*

$$\Sigma_B = \frac{1}{b}\mathbb{E}[x_{B,r}x_{B,r}^\top \mid H_B], \tag{157}$$

where H_B contains all preceding-layer quantities. Suppose there exist $\alpha_B > 0$, $\varepsilon_{\mathrm{diag},B} \geq 0$, and $\varepsilon_{\mathrm{off},B} \geq 0$ such that

$$|(\Sigma_B)_{ii} - \alpha_B| \leq \alpha_B \varepsilon_{\mathrm{diag},B}, \qquad i \in B, \tag{158}$$

$$\sum_{j\in B: j\neq i} |(\Sigma_B)_{ij}| \leq \alpha_B \varepsilon_{\mathrm{off},B}, \qquad i \in B. \tag{159}$$

Then every eigenvalue of Σ_B lies in

$$\alpha_B[1-\delta_{\infty,B}, 1+\delta_{\infty,B}], \qquad \delta_{\infty,B} = \varepsilon_{\mathrm{diag},B} + \varepsilon_{\mathrm{off},B}. \tag{160}$$

In particular, if $\max_{i\neq j}|(\Sigma_B)_{ij}| \leq \alpha_B \mu_B$, then one may take $\varepsilon_{\mathrm{off},B} \leq (b-1)\mu_B$.

Proof. For row i, Gershgorin's theorem places every eigenvalue in a disk centered at $(\Sigma_B)_{ii}$ with radius $\sum_{j\neq i}|(\Sigma_B)_{ij}|$. By Equations (158)– (159), every such disk is contained in

$$[\alpha_B(1-\varepsilon_{\mathrm{diag},B}-\varepsilon_{\mathrm{off},B}), \alpha_B(1+\varepsilon_{\mathrm{diag},B}+\varepsilon_{\mathrm{off},B})]. \tag{161}$$

Because Σ_B is symmetric, its eigenvalues are real, proving Equation (160). The coherence statement follows from summing at most $b-1$ off-diagonal entries in a row.

For each concrete GNN, Σ_B is the conditional graph-channel covariance created by its standard final random layer. Thus the mathematical content is not that all GCN or GIN batches are automatically isotropic. It is that comparable per-graph channel variance together with small aggregate cross-graph channel covariance implies the required mean spectral margin. Individual graphs may have larger or smaller channel values; only the batch covariance is constrained. □

Lemma B.24 (Matrix-Chernoff concentration of the complete batch readout)**.** *Fix a graph set B of size b and condition on H_B. Define*

$$Q_B = \frac{1}{mb}\sum_{r=1}^{m} x_{B,r}x_{B,r}^\top. \tag{162}$$

Suppose the vectors in Lemma B.22 are conditionally independent, the mean matrix $\Sigma_B = \mathbb{E}[Q_B \mid H_B]$ has eigenvalues in

$$[\alpha_B(1-\delta_{\infty,B}), \alpha_B(1+\delta_{\infty,B})], \tag{163}$$

and

$$\frac{\|x_{B,r}\|_2^2}{b} \leq L_B\alpha_B \qquad \textit{almost surely}. \tag{164}$$

Then, for every $\varepsilon \in (0,1)$, with conditional probability at least

$$1 - p_{\text{iso}}(B,\varepsilon), \qquad p_{\text{iso}}(B,\varepsilon) = 2b \exp\left[-\frac{m(1-\delta_{\infty,B})\varepsilon^2}{3L_B}\right], \tag{165}$$

all eigenvalues of Q_B lie in

$$[\alpha_B(1-\delta_{\infty,B})(1-\varepsilon), \alpha_B(1+\delta_{\infty,B})(1+\varepsilon)]. \tag{166}$$

Equivalently, the W_{out} parameter block satisfies batch scaled isometry with

$$\alpha_{B,m} = \alpha_B(1+\delta_{\infty,B}\varepsilon), \qquad \delta_{B,m} = \frac{\delta_{\infty,B}+\varepsilon}{1+\delta_{\infty,B}\varepsilon} \le \delta_{\infty,B}+\varepsilon. \tag{167}$$

Proof. Set

$$X_r = \frac{1}{mb} x_{B,r} x_{B,r}^\top. \tag{168}$$

Then $X_r \succeq 0$ and Equation (164) gives

$$\lambda_{\max}(X_r) = \frac{\|x_{B,r}\|_2^2}{mb} \le \frac{L_B \alpha_B}{m} =: R. \tag{169}$$

Moreover,

$$\sum_{r=1}^{m} \mathbb{E}[X_r \mid H_B] = \Sigma_B. \tag{170}$$

Hence

$$\mu_{\min} := \lambda_{\min}(\Sigma_B) \ge \alpha_B(1-\delta_{\infty,B}), \qquad \mu_{\max} := \lambda_{\max}(\Sigma_B) \ge \mu_{\min}. \tag{171}$$

Tropp's matrix Chernoff inequality (Tropp, 2012, Theorem 1.1, PDF p. 3) gives, for $\varepsilon \in (0,1)$,

$$\mathbb{P}[\lambda_{\min}(Q_B) \le (1-\varepsilon)\mu_{\min} | H_B] \le b \exp\left[-\frac{\mu_{\min}\varepsilon^2}{2R}\right], \tag{172}$$

$$\mathbb{P}[\lambda_{\max}(Q_B) \ge (1+\varepsilon)\mu_{\max} | H_B] \le b \exp\left[-\frac{\mu_{\max}\varepsilon^2}{3R}\right]. \tag{173}$$

Both exponents are at least $m(1-\delta_{\infty,B})\varepsilon^2/(3L_B)$ because $\mu_{\max} \ge \mu_{\min} \ge \alpha_B(1-\delta_{\infty,B})$. The upper endpoint in Equation (166) is at least $(1+\varepsilon)\mu_{\max}$, and the lower endpoint is at most $(1-\varepsilon)\mu_{\min}$. A union bound therefore proves Equations (165)– (166).

Let the lower and upper endpoints be m'_B and M'_B. Direct calculation gives

$$\frac{M'_B + m'_B}{2} = \alpha_B(1+\delta_{\infty,B}\varepsilon), \tag{174}$$

$$\frac{M'_B - m'_B}{M'_B + m'_B} = \frac{\delta_{\infty,B}+\varepsilon}{1+\delta_{\infty,B}\varepsilon}, \tag{175}$$

which proves Equation (167).

Each X_r is rank one and may be highly anisotropic. Thus neither an individual channel nor an individual graph must satisfy an isometry. The concentration occurs only after summing all width channels for the complete batch. □

Theorem B.25 (High-probability batch isometry of concrete GCN, GIN, and MPNN students)**.** *Consider any architecture in Definition B.20. Suppose its standard final width-m layer satisfies Lemma B.22, and suppose Lemmas B.23 and B.24 hold for a graph set B. Then the ordinary graph-projection block W_{out} satisfies Definition B.2 with probability at least $1 - p_{\text{iso}}(B,\varepsilon)$ and distortion $\delta_{B,m}$.*

For end-to-end training of all GNN parameters, let $\mathcal{K}_{\text{trunk},B}$ be the stacked tangent Gram of the message-passing trunk. If, on the actual routing plane,

$$0 \preceq \boldsymbol{U}_B^\top \mathcal{K}_{\text{trunk},B} \boldsymbol{U}_B \preceq \rho_B \alpha_{B,m} I, \tag{176}$$

then the complete GNN satisfies a batch restricted scaled isometry with

$$\delta_{\mathrm{GNN},B} \le \frac{2\delta_{B,m} + \rho_B}{2 + \rho_B}. \tag{177}$$

If the trunk itself satisfies a restricted scaled isometry with $(\alpha_{\mathrm{trunk},B}, \delta_{\mathrm{trunk},B})$*, then*

$$\delta_{\mathrm{GNN},B} \le \frac{\alpha_{B,m}\delta_{B,m} + \alpha_{\mathrm{trunk},B}\delta_{\mathrm{trunk},B}}{\alpha_{B,m} + \alpha_{\mathrm{trunk},B}}. \tag{178}$$

Proof. For GCN, GIN, and normalized MPNN, the standard final-layer construction in Lemma B.22 supplies the independent channel vectors. Proposition B.28, applied to $z_i/\sqrt{m}$, shows that the tangent Gram of W_{out} is $Q_B \otimes I_d$. Lemma B.24 controls all eigenvalues of Q_B, so the resulting event is uniform over every routing plane, including the plane generated adaptively by FORTE from the same network. No independence between the routing plane and Q_B is needed.

Let the output-block restricted eigenvalues be

$$m_{\mathrm{out}} = \alpha_{B,m}(1 - \delta_{B,m}), \qquad M_{\mathrm{out}} = \alpha_{B,m}(1 + \delta_{B,m}). \tag{179}$$

Under Equation (176), positivity of the trunk Gram gives

$$m_{\mathrm{total}} \ge m_{\mathrm{out}}, \qquad M_{\mathrm{total}} \le M_{\mathrm{out}} + \rho_B \alpha_{B,m}. \tag{180}$$

Therefore

$$\delta_{\mathrm{GNN},B} = \frac{M_{\mathrm{total}} - m_{\mathrm{total}}}{M_{\mathrm{total}} + m_{\mathrm{total}}} \tag{181}$$

$$\le \frac{\alpha_{B,m}(1 + \delta_{B,m} + \rho_B) - \alpha_{B,m}(1 - \delta_{B,m})}{\alpha_{B,m}(1 + \delta_{B,m} + \rho_B) + \alpha_{B,m}(1 - \delta_{B,m})} \tag{182}$$

$$= \frac{2\delta_{B,m} + \rho_B}{2 + \rho_B}. \tag{183}$$

If the trunk has its own two-sided interval, add the output and trunk lower bounds and add their upper bounds. The midpoint scales add, while the normalized half-width is their scale-weighted average, proving Equation (178).

This theorem includes three concrete GNN training regimes. If the GCN, GIN, or MPNN trunk is frozen and only W_{out} and the downstream heads are trained, then $\rho_B = 0$ and the readout probability applies directly. If the whole GNN is trained but its trunk tangent scale is dominated by the nearly isometric graph-projection block, Equation (176) applies. If both the trunk and graph projection are wide and each has a well-conditioned limiting tangent metric, the weighted bound applies. The GNTK construction of Du et al. (2019, Section 3.1, paper pp. 3–4) identifies deterministic infinite-width tangent kernels for message-passing GNNs; the additional requirement here is that the corresponding kernel be well conditioned on the current batch routing plane. Positive semidefiniteness alone is insufficient. □

Corollary B.26 (Joint full-data, mini-batch, and direction-concentration probability)**.** *Let* $B_0, \dots, B_{T-1}$ *be mini-batches sampled independently of GNN initialization. For the full graph set* $[N]$ *and the realized mini-batches, define*

$$p_N = 2N \exp\left[-\frac{m(1 - \delta_{\infty,N})\varepsilon_N^2}{3L_N}\right], \tag{184}$$

$$p_{B_t} = 2|B_t| \exp\left[-\frac{m(1 - \delta_{\infty,B_t})\varepsilon_b^2}{3L_{B_t}}\right]. \tag{185}$$

Before finite-horizon Jacobian drift,

$$\mathbb{P}\left(\mathcal{E}_{[N]} \cap \bigcap_{t=0}^{T-1} \mathcal{E}_{B_t}\right) \ge 1 - p_N - \sum_{t=0}^{T-1} p_{B_t}. \tag{186}$$

A sufficient width for the right-hand failure probability to be at most δ_{iso} is

$$m \geq \max\left\{ \frac{3L_N}{(1-\delta_{\infty,N})\varepsilon_N^2} \log \frac{4N}{\delta_{\mathrm{iso}}}, \max_{t<T} \frac{3L_{B_t}}{(1-\delta_{\infty,B_t})\varepsilon_b^2} \log \frac{4T|B_t|}{\delta_{\mathrm{iso}}} \right\}. \tag{187}$$

If Proposition B.33 keeps the resulting finite-horizon distortions below one, then:

(i) the full-data rate in Theorem B.7 holds with probability at least $1 - p_N$;

(ii) the mini-batch convergence bound for the fixed full empirical objective in Theorem B.41 holds with probability at least $1 - p_N - \delta_{\mathrm{mb}}$;

(iii) if one additionally wants every realized-batch descent certificate in Theorem B.5, then all full-data, mini-batch-isometry, and direction-concentration statements hold jointly with probability at least

$$1 - p_N - \sum_{t=0}^{T-1} p_{B_t} - \delta_{\mathrm{mb}}. \tag{188}$$

Proof. Condition on the realized mini-batch sequence. For each fixed graph set, Lemma B.24 gives the displayed failure probability. A union bound over $[N], B_0, \ldots, B_{T-1}$ gives Equation (186). Since the batch sequence is independent of initialization, integrating the conditional probability preserves the same bound. Splitting $\delta_{\mathrm{iso}}/2$ between the full set and the T mini-batches and solving each exponential inequality for m gives Equation (187).

The full-data theorem only needs $\mathcal{E}_{[N]}$. The stochastic full-objective theorem needs the same full-data oracle geometry plus its mini-batch direction-concentration event; it does not need each realized batch to be isometric. The per-batch descent theorem additionally needs all $\mathcal{E}_{B_t}$. Applying the corresponding union bounds proves the three probability statements. □

Corollary B.27 (Why batch-level matching permits individual graph fluctuations)**.** *Under the conditions of Lemma B.24, each rank-one channel matrix $x_{B,r}x_{B,r}^\top/(mb)$ may be singular and maximally anisotropic, and an individual graph may have a channel norm far from the batch mean. Nevertheless the complete batch Gram is isometric with failure probability $p_{\mathrm{iso}}(B,\varepsilon)$.*

Under exact mean isotropy, a sufficient width has the order

$$m = \mathcal{O}\left(\frac{L_B}{\varepsilon^2} \log \frac{|B|T}{\delta} \right). \tag{189}$$

Thus the theory does not require pointwise per-graph isometry. A mini-batch event is typically easier than a uniform full-data event because it involves $|B| \ll N$ graph coordinates. The theorem uses a stronger all-directions spectral event so that it remains valid for the routing plane generated adaptively by the same network.

Proof. The first statement follows directly from the rank-one form of each summand and the concentration of their sum. Solving $p_{\mathrm{iso}}(B,\varepsilon) \leq \delta/T$ for m gives Equation (189). No corresponding bound is imposed on an individual graph or channel. Replacing B by $[N]$ yields the stronger full-data width requirement. Finally, controlling the complete spectrum of Q_B makes the event uniform over every two-dimensional subspace; hence it also covers the data- and gradient-dependent routing plane without a separate independence assumption. □

Proposition B.28 (Exact Gram formula for a GNN graph-readout parameter block)**.** *Let an arbitrary message-passing GNN produce graph embeddings $z_i \in \mathbb{R}^m$, and let a shared trainable graph-level linear map produce the routed representation $\boldsymbol{h}_i = W z_i$. For a batch B of size b, let $Z_B \in \mathbb{R}^{b\times m}$ have rows $z_i^\top$, and define*

$$Q_B = \frac{1}{b} Z_B Z_B^\top. \tag{190}$$

Then the Jacobian Gram contributed by the parameter block W, acting on the stacked representation space, is

$$\mathcal{K}_{W,B} = Q_B \otimes I_d. \tag{191}$$

Consequently, if $\|Q_B - \alpha_B I_b\|_{\mathrm{op}} \le \alpha_B \delta_B$, then this GNN parameter block satisfies Definition B.2 on every routing plane, not only on the particular plane $\mathcal{S}_B$.

Proof. Let $v = (v_i)_{i\in B}$ be an arbitrary stacked representation vector and let $V \in \mathbb{R}^{b\times d}$ have rows $v_i^\top$. A perturbation ΔW changes graph i's routed representation by $\Delta \boldsymbol{h}_i = \Delta W z_i$. Therefore the adjoint Jacobian action is

$$\mathcal{J}_{W,B}^\top \frac{1}{\sqrt{b}}(v_i)_{i\in B} = \frac{1}{b}\sum_{i\in B} v_i z_i^\top = \frac{1}{b} V^\top Z_B. \tag{192}$$

Its squared Frobenius norm is

$$\left\| \frac{1}{b} V^\top Z_B \right\|_F^2 = \frac{1}{b^2}\operatorname{tr}(V^\top Z_B Z_B^\top V) \tag{193}$$

$$= \frac{1}{b}\operatorname{tr}(V^\top Q_B V). \tag{194}$$

Meanwhile,

$$\left\| \frac{1}{\sqrt{b}}(v_i)_{i\in B} \right\|_2^2 = \frac{1}{b}\operatorname{tr}(V^\top V). \tag{195}$$

Thus the quadratic form of the stacked Jacobian Gram is exactly the quadratic form of $Q_B \otimes I_d$, proving Equation (191). The final claim follows directly from the operator-norm inequality.

This formula yields concrete GNN cases inside the proof. If the pooled GNN embeddings have equal norm and are mutually orthogonal on B, then $Q_B = \alpha_B I_b$ and $\delta_B = 0$. More generally, suppose

$$\left| \frac{\|z_i\|_2^2}{b} - \alpha_B \right| \le \alpha_B \varepsilon_{\mathrm{diag}}, \qquad \frac{1}{b}\sum_{j\in B: j\neq i} |z_i^\top z_j| \le \alpha_B \varepsilon_{\mathrm{off}} \tag{196}$$

for every $i \in B$. Gershgorin's theorem places every eigenvalue of Q_B in

$$\alpha_B[1 - (\varepsilon_{\mathrm{diag}} + \varepsilon_{\mathrm{off}}), 1 + (\varepsilon_{\mathrm{diag}} + \varepsilon_{\mathrm{off}})]. \tag{197}$$

Hence Definition B.2 holds with $\delta_B \le \varepsilon_{\mathrm{diag}} + \varepsilon_{\mathrm{off}}$. This is a checkable condition on graph-level GNN embeddings; it does not follow merely from using mean pooling or normalization. □

Corollary B.29 (Exact restricted criterion for the GNN readout block)**.** *Let $\boldsymbol{U}_B$ have orthonormal columns spanning the actual batch routing plane $\mathcal{S}_B$. For the graph-readout block in Proposition B.28, form the at-most-2×2 matrix*

$$H_{W,B} = \boldsymbol{U}_B^\top (Q_B \otimes I_d)\boldsymbol{U}_B. \tag{198}$$

If $m_{W,B} = \lambda_{\min}(H_{W,B}) > 0$ and $M_{W,B} = \lambda_{\max}(H_{W,B})$, then the block satisfies Definition B.2 on the actual routing plane with

$$\alpha_{W,B} = \frac{M_{W,B} + m_{W,B}}{2}, \qquad \delta_{W,B} = \frac{M_{W,B} - m_{W,B}}{M_{W,B} + m_{W,B}}. \tag{199}$$

This conclusion does not require Q_B to be nonsingular on the entire b-dimensional graph-index space. Thus it remains applicable when the number of graphs exceeds the embedding width, provided the two routed gradient directions avoid the null directions of the readout Jacobian.

Proof. By Proposition B.28, the readout block's stacked Jacobian Gram is $Q_B \otimes I_d$. Compressing this operator to $\mathcal{S}_B$ gives Equation (198). Every $v \in \mathcal{S}_B$ can be written as $v = U_B c$, and hence

$$v^\top (Q_B \otimes I_d) v = c^\top H_{W,B} c. \tag{200}$$

The Rayleigh quotient of $H_{W,B}$ lies in $[m_{W,B}, M_{W,B}]$. Rewriting these endpoints as $\alpha_{W,B}(1 \mp \delta_{W,B})$ proves the claimed restricted scaled isometry. Positivity of $m_{W,B}$ is exactly the statement that neither of the two relevant routed directions is annihilated by this parameter block. □

Lemma B.30 (Aggregation of GNN parameter blocks)**.** *Suppose the GNN parameters are partitioned into blocks and the corresponding batch Jacobian Gram is $\mathcal{K}_B = \sum_{\ell=1}^{L} \mathcal{K}_{B,\ell}$. If block ℓ obeys*

$$\alpha_{B,\ell}(1-\delta_{B,\ell})\|v\|_2^2 \le v^\top \mathcal{K}_{B,\ell} v \le \alpha_{B,\ell}(1+\delta_{B,\ell})\|v\|_2^2 \tag{201}$$

for every $v \in \mathcal{S}_B$, then the complete GNN Jacobian satisfies Definition B.2 with

$$\alpha_B = \sum_{\ell=1}^{L} \alpha_{B,\ell}, \qquad \delta_B \le \frac{\sum_{\ell=1}^{L} \alpha_{B,\ell}\delta_{B,\ell}}{\sum_{\ell=1}^{L} \alpha_{B,\ell}}. \tag{202}$$

Proof. Summing the lower bounds in Equation (201) gives

$$v^\top \mathcal{K}_B v \ge \sum_{\ell} \alpha_{B,\ell}(1-\delta_{B,\ell})\|v\|_2^2 \tag{203}$$

$$= \alpha_B \left(1 - \frac{\sum_\ell \alpha_{B,\ell}\delta_{B,\ell}}{\alpha_B}\right)\|v\|_2^2. \tag{204}$$

The upper bound is identical with the signs reversed. Therefore a nearly isometric graph-readout block from Proposition B.28 can stabilize the complete GNN metric when it carries a substantial fraction of the total tangent-feature scale, but an arbitrary highly anisotropic trunk is not automatically harmless. □

Proposition B.31 (High-probability isometry for a wide-channel GNN block)**.** *Fix a graph set B and consider a width-m GNN parameter block whose stacked Jacobian Gram admits the channel decomposition*

$$\mathcal{K}_{B,0} = \frac{1}{m}\sum_{r=1}^{m} Y_{B,r}, \tag{205}$$

where, conditional on the preceding GNN layers, the self-adjoint matrices $Y_{B,r}$ are independent with common mean $\overline{\mathcal{K}}_B$. Let

$$X_{B,r} = Y_{B,r} - \overline{\mathcal{K}}_B, \quad \|X_{B,r}\|_{\mathrm{op}} \le L_B, \quad V_B = \left\|\mathbb{E}[X_{B,r}^2]\right\|_{\mathrm{op}}, \quad \kappa_B = \frac{\mathrm{tr}(\mathbb{E}[X_{B,r}^2])}{V_B}. \tag{206}$$

Suppose the infinite-width mean metric satisfies

$$\|\overline{\mathcal{K}}_B - \alpha_B I\|_{\mathrm{op}} \le \alpha_B \delta_{\infty,B}. \tag{207}$$

Then, for every $s > 0$, with probability at least

$$1 - 2\kappa_B \frac{s}{e^s - s - 1}, \tag{208}$$

we have

$$\|\mathcal{K}_{B,0} - \overline{\mathcal{K}}_B\|_{\mathrm{op}} \le \varepsilon_{B,m}(s) := \sqrt{\frac{2V_B s}{m}} + \frac{L_B s}{3m}, \tag{209}$$

and therefore

$$\delta_{B,0} \le \delta_{\infty,B} + \frac{\varepsilon_{B,m}(s)}{\alpha_B}, \tag{210}$$

provided the right-hand side is smaller than one.

Proof. Apply the intrinsic-dimension matrix Bernstein inequality of Hsu, Kakade, and Zhang (Hsu et al., 2012, Theorem 3.3, paper pp. 4–5) to $m^{-1}\sum_r X_{B,r}$ and to its negative. Each one-sided event has probability at most $\kappa_B s/(e^s - s - 1)$ at the radius in Equation (209); a union bound gives Equations (208)– (209). By the triangle inequality,

$$\|\mathcal{K}_{B,0} - \alpha_B I\|_{\mathrm{op}} \le \|\mathcal{K}_{B,0} - \overline{\mathcal{K}}_B\|_{\mathrm{op}} + \|\overline{\mathcal{K}}_B - \alpha_B I\|_{\mathrm{op}} \tag{211}$$

$$\le \varepsilon_{B,m}(s) + \alpha_B \delta_{\infty,B}. \tag{212}$$

This is exactly Definition B.2 with the distortion in Equation (210).

A standard message-passing GNN instance of Equation (205) arises as follows. Let a normalized message-passing trunk produce the width-normalized graph embedding

$$z_i = \frac{1}{\sqrt{m}}(\xi_{i,1}, \dots, \xi_{i,m}) \in \mathbb{R}^m, \tag{213}$$

and train a graph-level linear map. For channel r, set $x_{B,r} = (\xi_{i,r})_{i \in B}$. Proposition B.28 then gives

$$Q_B = \frac{1}{b} Z_B Z_B^\top = \frac{1}{m} \sum_{r=1}^{m} \frac{1}{b} x_{B,r} x_{B,r}^\top, \tag{214}$$

$$\mathcal{K}_{B,0} = \frac{1}{m} \sum_{r=1}^{m} Y_{B,r}, \qquad Y_{B,r} = \frac{1}{b} x_{B,r} x_{B,r}^\top \otimes I_d. \tag{215}$$

Conditional on the preceding layers, independently initialized channels give the independence required in the proposition. Bounded graph size, normalized aggregation, bounded node features, and a bounded activation imply finite L_B; the variance quantity V_B records the actual channel variability.

The mean-isotropy condition also has a concrete GNN channel interpretation. If, conditionally on the preceding layers,

$$\mathbb{E}[x_{B,r} x_{B,r}^\top] = b \alpha_B I_b, \tag{216}$$

then $\overline{\mathcal{K}}_B = \alpha_B I$ and $\delta_{\infty,B} = 0$. More generally, if every diagonal entry of $b^{-1}\mathbb{E}[x_{B,r} x_{B,r}^\top]$ differs from α_B by at most $\alpha_B \varepsilon_{\mathrm{diag}}$ and every row's sum of absolute off-diagonal entries is at most $\alpha_B \varepsilon_{\mathrm{off}}$, Gershgorin's theorem gives

$$\delta_{\infty,B} \le \varepsilon_{\mathrm{diag}} + \varepsilon_{\mathrm{off}}. \tag{217}$$

Thus the mathematical requirement is equal channel variance together with small cross-graph channel covariance; width then concentrates the finite GNN metric around that mean. Width alone does not force the mean to be isotropic. The GNTK construction of Du et al. (Du et al., 2019, Sections 2–3, paper pp. 2–5) gives the deterministic infinite-width tangent kernel corresponding to message-passing architectures, including GCN- and GIN-style aggregation. In the present theorem, that kernel must additionally satisfy the restricted spectral condition in Equation (207); positive definiteness alone is not enough. □

Corollary B.32 (Simultaneous full-data and mini-batch initialization event)**.** *Condition on a sequence of T sampled mini-batches $B_0, \dots, B_{T-1}$ that is independent of the GNN initialization. Apply Proposition B.31 to the $T+1$ graph sets $[N], B_0, \dots, B_{T-1}$. Let $\kappa_{\max}$ bound their intrinsic dimensions and set*

$$s_\delta = \max\left\{2.6, 2\log\frac{2(T+1)\kappa_{\max}}{\delta_{\mathrm{iso}}}\right\}. \tag{218}$$

Then, with probability at least $1 - \delta_{\mathrm{iso}}$ over initialization, Equation (210) holds simultaneously for the full dataset and all T realized mini-batches, with $s = s_\delta$ and their own $(\alpha_B, V_B, L_B, \delta_{\infty,B})$.

Proof. For $s \ge 2.6$, Hsu et al. show $s/(e^s - s - 1) \le e^{-s/2}$ (Hsu et al., 2012, p. 2). Therefore each graph set fails with probability at most $2\kappa_{\max} e^{-s_\delta/2} \le \delta_{\mathrm{iso}}/(T+1)$. A union bound over the full dataset and the T batches proves the result. Independence between the batch sequence and initialization permits conditioning on the realized sets before applying the matrix concentration inequality. □

Proposition B.33 (Persistence of a GNN batch isometry during training)**.** *Fix a graph set B. Suppose*

$$\|\boldsymbol{J}_B(w) - \boldsymbol{J}_B(w_0)\|_{\mathrm{op}} \le L_{J,B} \|w - w_0\|_2 \tag{219}$$

on a parameter ball of radius $\mathcal{R}$. If the initialization satisfies

$$\|\mathcal{K}_{B,0} - \alpha_B I\|_{\mathrm{op}} \le \alpha_B \delta_{B,0}, \tag{220}$$

and $\|w_t - w_0\|_2 \le \mathcal{R}$*, then*

$$\delta_{B,t} \le \delta_{B,0} + \frac{(2\|\mathcal{J}_{B,0}\|_{\mathrm{op}} + L_{J,B}\mathcal{R})L_{J,B}\mathcal{R}}{\alpha_B}. \tag{221}$$

Proof. Write $\Delta \boldsymbol{J}_B = \mathcal{J}_{B,t} - \mathcal{J}_{B,0}$. Since $\mathcal{K}_{B,t} = \mathcal{J}_{B,t}\mathcal{J}_{B,t}^\top$,

$$\mathcal{K}_{B,t} - \mathcal{K}_{B,0} = \Delta \boldsymbol{J}_B \mathcal{J}_{B,t}^\top + \mathcal{J}_{B,0}\Delta \boldsymbol{J}_B^\top. \tag{222}$$

Hence

$$\|\mathcal{K}_{B,t} - \mathcal{K}_{B,0}\|_{\mathrm{op}} \le (\|\mathcal{J}_{B,t}\|_{\mathrm{op}} + \|\mathcal{J}_{B,0}\|_{\mathrm{op}})\|\Delta \boldsymbol{J}_B\|_{\mathrm{op}} \tag{223}$$

$$\le (2\|\mathcal{J}_{B,0}\|_{\mathrm{op}} + L_{J,B}\mathcal{R})L_{J,B}\mathcal{R}. \tag{224}$$

Combining this bound with Equation (220) by the triangle inequality proves Equation (221).

The Lipschitz condition in Equation (219) has a concrete GNN sufficient case. Consider a finite-depth message-passing network

$$H_i^{(0)} = X_i, \qquad H_i^{(\ell)} = \sigma_\ell(\widetilde{A}_i H_i^{(\ell-1)} W_\ell), \qquad \boldsymbol{h}_i = P_i H_i^{(L)}, \tag{225}$$

where $\|\widetilde{A}_i\|_{\mathrm{op}} \le 1$, graph size and $\|X_i\|_F$ are bounded, $\|P_i\|_{\mathrm{op}}$ is bounded, the weight spectral norms remain bounded on the analyzed ball, and each σ_ℓ has bounded first and second derivatives. Repeated chain rule gives a finite bound on the output Hessian $\|D_w^2 \boldsymbol{h}_i(w)\|_{\mathrm{op}}$ over that ball. The mean-value theorem then gives

$$\|D_w h_i(w) - D_w h_i(w_0)\|_{\mathrm{op}} \le \sup_{\widetilde{w} \text{ on the segment}} \|D_w^2 \boldsymbol{h}_i(\widetilde{w})\|_{\mathrm{op}} \|w - w_0\|_2. \tag{226}$$

Stacking the finitely many graphs in B yields Equation (219). Thus a smooth, normalized, finite-depth GNN in a bounded parameter neighborhood has a finite $L_{J,B}$. The quantitative requirement for persistence is that the drift term in Equation (221) remain smaller than the remaining isometry margin; merely calling an architecture "residual" or "wide" does not establish this fact. □

Corollary B.34 (A finite-horizon persistence condition for bounded GNN updates)**.** *Under Proposition B.33, suppose the actual parameter directions satisfy* $\|\widehat{d}_t\|_2 \le B_d$ *for* $0 \le t < T$*. Then*

$$\|w_t - w_0\|_2 \le \eta \sum_{s=0}^{t-1} \|\widehat{d}_s\|_2 \le \eta T B_d. \tag{227}$$

Consequently, for every graph set B *whose initialization event holds,*

$$\delta_{B,t} \le \delta_{B,0} + \frac{(2\|\mathcal{J}_{B,0}\|_{\mathrm{op}} + L_{J,B}\eta T B_d)L_{J,B}\eta T B_d}{\alpha_B}. \tag{228}$$

In particular, if the right-hand side is at most a prescribed $\bar{\delta}_B < 1$*, then the batch restricted isometry persists throughout the entire horizon. Proposition B.47 below supplies a concrete finite value of* B_d *for bounded normalized message-passing GNNs.*

Proof. The first inequality is the triangle inequality applied to $w_t - w_0 = -\eta \sum_{s<t} \widehat{d}_s$. The direction bound gives the second. Substitute $\mathcal{R} = \eta T B_d$ into Equation (221). If the resulting upper bound is less than one, both restricted eigenvalues remain positive, proving persistence. □

Corollary B.35 (Explicit high-probability full-data GNN convergence)**.** *Apply Proposition B.31 to* $B = [N]$ *with confidence parameter* s*, and suppose the finite-horizon drift term in Equation (228) leaves a final distortion* $\bar{\delta}_{[N]} < 1$*. Form* $c_\star, D_\star$ *from Equations (47)–(48) using this worst-case distortion and suppose* $c_\star > 0$*. Then the full-data rate*

$$\min_{0 \le t < T} \|\nabla F_N(w_t)\|_2^2 \le \frac{2\beta D_\star^2 [F_N(w_0) - F_\star]}{c_\star^2 T} \tag{229}$$

holds with probability at least

$$1 - 2\kappa_{[N]} \frac{s}{e^s - s - 1} \tag{230}$$

over the random GNN channels. Thus the full-data theorem is genuinely high-probability: concentration establishes the initialization metric, and the finite-horizon drift bound propagates it to every iterate.

Proof. Proposition B.31 gives the initialization restricted metric with the probability in Equation (230). On that event, Corollary B.34 gives the required isometry for every $0 \le t < T$. Theorem B.7 then gives Equation (229). □

B.6 Actual mini-batch direction and high-probability convergence

At iteration t, condition on the actual pre-batch parameters and the actual pre-batch available-only EMA state. Let B_t be a random mini-batch of size b_t. Starting from this same state, define a one-step full-data oracle EMA update by replacing each current mini-batch available-sample mean with its full-data available-sample mean. Let d_t be the average of the resulting per-graph routed parameter directions over all N graphs. Let $\widehat{d}_t$ be the direction produced by the unchanged algorithm from B_t, including its actual current-batch available-only EMA update. The actual update is

$$w_{t+1} = w_t - \eta \widehat{d}_t. \tag{231}$$

The stochastic analysis separates ordinary graph subsampling from the extra perturbation introduced because the current mini-batch is also used to update the EMA scale before routing.

Definition B.36 (Full-data oracle scale and oracle-scale mini-batch)**.** For teacher k, let $\nu^{\circ}_{k,t}$ be the one-step bias-corrected EMA obtained from the actual pre-batch state by inserting the full-data available-sample mean of $\|\boldsymbol{u}_{k,i}\|_2$. This counterfactual scale is defined whether or not teacher k happens to occur in the sampled batch. Let $\psi_{i,t}(\nu^{\circ}_t)$ denote graph i's complete deployed-parameter direction under these fixed scales, including its availability mask and the unchanged router. Define

$$d_t = \frac{1}{N} \sum_{i=1}^{N} \psi_{i,t}(\nu^{\circ}_t), \qquad \widetilde{d}_t = \frac{1}{b_t} \sum_{i \in B_t} \psi_{i,t}(\nu^{\circ}_t). \tag{232}$$

Because the same scale vector is held fixed in both averages, $\widetilde{d}_t - d_t$ is an ordinary vector-valued graph-sampling error. The actual direction admits the exact decomposition

$$\widehat{d}_t - d_t = \underbrace{(\widetilde{d}_t - d_t)}_{\text{graph-sampling error}} + \underbrace{(\widehat{d}_t - \widetilde{d}_t)}_{\text{current-EMA error}}. \tag{233}$$

If a teacher is absent from B_t, changing its scale does not change either mini-batch direction because all of its availability masks are zero.

Lemma B.37 (Exact coefficient of the current batch in the available-only EMA)**.** *Suppose teacher k has been observed in $n^{-}_{k,t}$ previous batches. If it is observed in the current batch, define*

$$\alpha^{\mathrm{EMA}}_{k,t} = \frac{1 - \rho}{1 - \rho^{n^{-}_{k,t}+1}}. \tag{234}$$

Let $y^{B}_{k,t}$ and $y^{N}_{k,t}$ be the current mini-batch and full-data available-sample means of $\|\boldsymbol{u}_{k,i}\|_2$. Then

$$\widehat{\nu}_{k,t} - \nu^{\circ}_{k,t} = \alpha^{\mathrm{EMA}}_{k,t} (y^{B}_{k,t} - y^{N}_{k,t}). \tag{235}$$

If the teacher is absent from the entire batch, the actual EMA is not updated and Equation (235) is not invoked; the teacher's contribution to $\widehat{d}_t - \widetilde{d}_t$ is exactly zero because all of its mini-batch availability masks vanish.

Proof. Let $v^-_{k,t}$ be the uncorrected pre-batch EMA. On an observed batch, the actual corrected update is

$$\widehat{\nu}_{k,t} = \frac{\rho v^-_{k,t} + (1-\rho) y^B_{k,t}}{1 - \rho^{n^-_{k,t}+1}}, \tag{236}$$

whereas the oracle scale is the same expression with $y^B_{k,t}$ replaced by $y^N_{k,t}$. Subtraction proves Equation (235). When the teacher is absent, its scale may differ from the counterfactual full-data oracle scale, but this difference is multiplied by zero availability on every sampled graph and therefore creates no current-direction error. At the next iteration the oracle comparison is rebuilt from the actual inherited EMA state, so no untracked oracle state is propagated across time. □

Lemma B.38 (Lipschitz propagation of an EMA-scale perturbation to GNN parameters)**.** *Fix a graph i and all task and teacher representation gradients. Let $\boldsymbol{a}_i(\nu)$ denote the unchanged routed auxiliary representation gradient when the vector of EMA scales is ν. Then*

$$\|\boldsymbol{a}_i(\nu) - \boldsymbol{a}_i(\nu')\|_2 \le \frac{\|\boldsymbol{g}_i\|_2}{4\epsilon}\left(2 + \frac{\sqrt{K}}{\tau}\right)\|\nu - \nu'\|_2. \tag{237}$$

Consequently,

$$\|\boldsymbol{J}_i^\top[\boldsymbol{a}_i(\nu) - \boldsymbol{a}_i(\nu')]\|_2 \le L_{\theta,i}\|\nu - \nu'\|_2, \quad L_{\theta,i} = \|\boldsymbol{J}_i\|_{\mathrm{op}}\frac{\|\boldsymbol{g}_i\|_2}{4\epsilon}\left(2 + \frac{\sqrt{K}}{\tau}\right). \tag{238}$$

Proof. For teacher k, let $U_k = \|\boldsymbol{u}_{k,i}\|_2$ and $G = \|\boldsymbol{g}_i\|_2$. The relative magnitude entering the evidence is

$$q_k(\nu_k) = \min\left\{\frac{c_{\max}U_k}{G}, \frac{U_k}{\max\{U_k, \nu_k\} + \epsilon}\right\}. \tag{239}$$

The scalar map $x \mapsto \max\{U_k, x\}$ is one-Lipschitz and the outer minimum cannot increase a Lipschitz constant. On the branch $\nu_k > U_k$,

$$\left|\frac{d}{d\nu_k}\frac{U_k}{\nu_k + \epsilon}\right| = \frac{U_k}{(\nu_k + \epsilon)^2} \le \frac{U_k}{(U_k + \epsilon)^2} \le \frac{1}{4\epsilon}, \tag{240}$$

where the last step uses $(U_k + \epsilon)^2 \ge 4U_k\epsilon$. On the other branch the derivative is zero. Therefore

$$\|q(\nu) - q(\nu')\|_2 \le \frac{1}{4\epsilon}\|\nu - \nu'\|_2. \tag{241}$$

The cosine scores and availability mask do not depend on the EMA scale, so the evidence vector satisfies the same bound. The softmax Jacobian is $\tau^{-1}[\mathrm{Diag}(\omega) - \omega\omega^\top]$ and has operator norm at most $1/(2\tau)$, hence

$$\|\omega - \omega'\|_2 \le \frac{1}{2\tau}\|e - e'\|_2. \tag{242}$$

For $r = \omega^\top e$,

$$|r - r'| \le |\omega^\top(e - e')| + |(\omega - \omega')^\top e'| \tag{243}$$

$$\le \left(1 + \frac{\sqrt{K}}{2\tau}\right)\|e - e'\|_2. \tag{244}$$

Since $\pi = r\omega$,

$$\|\pi - \pi'\|_1 \le r\|\omega - \omega'\|_1 + |r - r'|\|\omega'\|_1 \tag{245}$$

$$\le \left(1 + \frac{\sqrt{K}}{\tau}\right)\|e - e'\|_2. \tag{246}$$

Finally, calibrated teacher gradients have norm at most G. Decomposing the difference of $a = \sum_k \pi_k \bar{u}_k$ into a change of the calibrated vectors and a change of the route masses gives

$$\|a(\nu) - a(\nu')\|_2 \le \frac{G}{4\epsilon}\|\nu - \nu'\|_2 + G\|\pi - \pi'\|_1 \tag{247}$$

$$\le \frac{G}{4\epsilon}\left(2 + \frac{\sqrt{K}}{\tau}\right)\|\nu - \nu'\|_2. \tag{248}$$

Applying $\|\boldsymbol{J}_i^\top v\|_2 \le \|\boldsymbol{J}_i\|_{\mathrm{op}}\|v\|_2$ proves the parameter-space statement. □

Definition B.39 (Conditional Bernstein radii)**.** Condition on the history before iteration t. The batch is sampled uniformly from the finite training set, either with replacement or by a simple random sample without replacement. For the latter, use the reveal filtration and its Doob martingale differences. Suppose the centered oracle-scale per-graph parameter directions in Equation (232) satisfy the Hilbert-space Bernstein moment condition with variance proxy $v_{d,t}$ and scale $\Gamma_{d,t}$. For teacher k, conditional on the realized available count $n_{k,t} > 0$, the sampled available graphs are uniform over the full-data available set. Suppose their centered teacher-norm observations satisfy the scalar Bernstein condition with parameters $(v_{\nu,k,t}, \Gamma_{\nu,k,t})$. For confidence parameter $\ell > 0$, define

$$\varepsilon_{d,t}(\ell) = \sqrt{\frac{2v_{d,t}\ell}{b_t}} + \frac{\Gamma_{d,t}\ell}{b_t}, \tag{249}$$

$$\varepsilon_{\nu,k,t}(\ell) = \alpha_{k,t}^{\mathrm{EMA}}\left[\sqrt{\frac{2v_{\nu,k,t}\ell}{n_{k,t}}} + \frac{\Gamma_{\nu,k,t}\ell}{n_{k,t}}\right], \tag{250}$$

where $n_{k,t}$ is the number of available teacher-k observations in the current batch. Conditional on $n_{k,t} > 0$, the radius controls the random available-sample mean. When $n_{k,t} = 0$, the second radius is defined as zero because that teacher contributes no route and hence no scale-induced current direction perturbation.

The dimension-free Hilbert-space Bernstein inequality used here is Theorem 3.3 of Pinelis (Pinelis, 1994, Theorem 3.3, pp. 1688–1689). It does not introduce the number of GNN parameters into the radius.

Lemma B.40 (One-step high-probability parameter-direction approximation)**.** *Let*

$$\overline{L}_{\theta,t} = \frac{\lambda}{b_t}\sum_{i\in B_t} L_{\theta,i}. \tag{251}$$

Under Definition B.39, with probability at least $1 - 2(K+1)e^{-\ell}$ *conditional on the pre-batch history,*

$$\|\widehat{d}_t - d_t\|_2 \le \varepsilon_t(\ell) := \varepsilon_{d,t}(\ell) + \overline{L}_{\theta,t}\left(\sum_{k=1}^{K}\varepsilon_{\nu,k,t}(\ell)^2\right)^{1/2}. \tag{252}$$

Proof. Pinelis' inequality applied to the centered oracle-scale parameter directions gives

$$\|\widetilde{d}_t - d_t\|_2 \le \varepsilon_{d,t}(\ell) \tag{253}$$

with conditional failure probability at most $2e^{-\ell}$. For each teacher, the scalar specialization and Lemma B.37 give

$$|\widehat{\nu}_{k,t} - \nu^\circ_{k,t}| \le \varepsilon_{\nu,k,t}(\ell) \tag{254}$$

with failure probability at most $2e^{-\ell}$. A union bound over the K scales and the direction event gives total conditional failure probability $2(K+1)e^{-\ell}$. On the intersection event, Lemma B.38 and the

triangle inequality give

$$\|\widehat{d}_t - \widetilde{d}_t\|_2 \le \frac{\lambda}{b_t} \sum_{i \in B_t} L_{\theta,i} \|\widehat{\nu}_t - \nu_t^\circ\|_2 \tag{255}$$

$$\le \overline{L}_{\theta,t} \left(\sum_k \varepsilon_{\nu,k,t}(\ell)^2 \right)^{1/2}. \tag{256}$$

Combining this with Equation (233) proves the claim. □

Theorem B.41 (High-probability mini-batch convergence of the full empirical objective)**.** *Suppose Assumption B.6 holds. On an initialization event $\mathcal{E}_{\mathrm{iso}}$, suppose the full-data quantities satisfy*

$$G_t^\top d_t \ge c_\star \|G_t\|_2^2, \qquad \|d_t\|_2 \le D_\star \|G_t\|_2 \tag{257}$$

for every $0 \le t < T$, with $c_\star > 0$. Let

$$0 < \eta < \frac{2c_\star}{\beta D_\star^2}, \qquad a_\star = c_\star - \frac{\beta \eta D_\star^2}{2} > 0, \qquad b_\star = 1 + \beta \eta D_\star. \tag{258}$$

Set

$$\ell_{T,\delta} = \log \frac{2T(K+1)}{\delta_{\mathrm{mb}}}. \tag{259}$$

Then, conditional on $\mathcal{E}_{\mathrm{iso}}$, with probability at least $1 - \delta_{\mathrm{mb}}$ over all mini-batch draws,

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F_N(w_t)\|_2^2 \le \frac{2[F_N(w_0) - F_\star]}{\eta a_\star T} + \frac{1}{a_\star T} \left(\frac{b_\star^2}{a_\star} + \beta \eta \right) \sum_{t=0}^{T-1} \varepsilon_t (\ell_{T,\delta})^2. \tag{260}$$

If $\mathbb{P}(\mathcal{E}_{\mathrm{iso}}) \ge 1 - \delta_{\mathrm{iso}}$, the joint probability of Equation (260) is at least $1 - \delta_{\mathrm{iso}} - \delta_{\mathrm{mb}}$.

Proof. For one iteration, write

$$\widehat{d}_t = d_t + e_t, \qquad \|e_t\|_2 \le \varepsilon_t. \tag{261}$$

By smoothness of the full empirical objective and the actual update Equation (231),

$$F_N(w_{t+1}) \le F_N(w_t) - \eta G_t^\top (d_t + e_t) + \frac{\beta \eta^2}{2} \|d_t + e_t\|_2^2 \tag{262}$$

$$\le F_N(w_t) - \eta c_\star \|G_t\|_2^2 + \eta \|G_t\|_2 \varepsilon_t \tag{263}$$

$$+ \frac{\beta \eta^2}{2} \left(D_\star^2 \|G_t\|_2^2 + 2 D_\star \|G_t\|_2 \varepsilon_t + \varepsilon_t^2 \right) \tag{264}$$

$$= F_N(w_t) - \eta a_\star \|G_t\|_2^2 + \eta b_\star \|G_t\|_2 \varepsilon_t + \frac{\beta \eta^2}{2} \varepsilon_t^2. \tag{265}$$

Young's inequality with parameter $a_\star$ gives

$$b_\star \|G_t\|_2 \varepsilon_t \le \frac{a_\star}{2} \|G_t\|_2^2 + \frac{b_\star^2}{2a_\star} \varepsilon_t^2. \tag{266}$$

Substitution into Equation (265) yields

$$F_N(w_{t+1}) \le F_N(w_t) - \frac{\eta a_\star}{2} \|G_t\|_2^2 + \frac{\eta}{2} \left(\frac{b_\star^2}{a_\star} + \beta \eta \right) \varepsilon_t^2. \tag{267}$$

It remains to justify that the direction bounds hold simultaneously. By Lemma B.40, conditional on the history at iteration t, the failure probability is at most $2(K+1)e^{-\ell_{T,\delta}} = \delta_{\mathrm{mb}}/T$. Taking unconditional probability uses the tower property:

$$\mathbb{P}(E_t^c) = \mathbb{E}[\mathbb{P}(E_t^c \mid \mathcal{F}_t)] \le \frac{\delta_{\mathrm{mb}}}{T}. \tag{268}$$

A union bound gives $\mathbb{P}(\cap_{t=0}^{T-1} E_t) \geq 1 - \delta_{\mathrm{mb}}$. On this intersection, Equation (267) holds for every t. Summing gives

$$\frac{\eta a_\star}{2} \sum_{t=0}^{T-1} \|G_t\|_2^2 \leq \sum_{t=0}^{T-1} [F_N(w_t) - F_N(w_{t+1})] \tag{269}$$

$$+ \frac{\eta}{2} \left(\frac{b_\star^2}{a_\star} + \beta\eta \right) \sum_{t=0}^{T-1} \varepsilon_t^2 \tag{270}$$

$$= F_N(w_0) - F_N(w_T) \tag{271}$$

$$+ \frac{\eta}{2} \left(\frac{b_\star^2}{a_\star} + \beta\eta \right) \sum_{t=0}^{T-1} \varepsilon_t^2 \tag{272}$$

$$\leq F_N(w_0) - F_\star \tag{273}$$

$$+ \frac{\eta}{2} \left(\frac{b_\star^2}{a_\star} + \beta\eta \right) \sum_{t=0}^{T-1} \varepsilon_t^2. \tag{274}$$

Division by $\eta a_\star T/2$ proves Equation (260). Finally, a union bound combines the initialization and mini-batch events. □

B.6.1 Mini-batch task-loss contraction with direction error

We extend Section B.4.1 using the existing one-step full-data oracle in Definition B.36. At the actual mini-batch iterate, let $d_t = G_t + \lambda A_t$ be that oracle direction and $\widehat{d}_t = d_t + e_t$ the actual direction. The error includes both graph sampling and the current-batch EMA update, as in Equation (233); no unbiasedness assumption is imposed on $\widehat{d}_t$.

Corollary B.42 (Task-loss contraction on a simultaneous error event)**.** *Suppose the smoothness and PL assumptions of Theorem B.16 hold. At every actual iterate, suppose the oracle direction satisfies the same gain–norm coupling with uniform $m > 0, \Lambda < \infty$ and $\|d_t\|_2 \leq D_{\mathrm{F}} \|G_t\|_2$. Here D_{F} is the trajectory bound in Equation (48), evaluated for the oracle directions along this stochastic trajectory. Use H_{F} and η from Theorem B.16. On a simultaneous event $\|e_t\|_2 \leq \varepsilon_t$ for all $t < T$, define*

$$C_\eta = \frac{\eta}{2} \left[\frac{(1 + \beta\eta D_{\mathrm{F}})^2}{H_{\mathrm{F}}} + \beta\eta \right], \qquad q_{\mathrm{mb}} = 1 - \mu\eta H_{\mathrm{F}} \in (0, 1). \tag{275}$$

Then

$$\Delta_T \leq q_{\mathrm{mb}}^T \Delta_0 + C_\eta \sum_{t=0}^{T-1} q_{\mathrm{mb}}^{T-1-t} \varepsilon_t^2. \tag{276}$$

The radius $\varepsilon_t = \varepsilon_t(\ell_{T,\delta})$ and the simultaneous probability event can be taken from Equation (252) and Theorem B.41, respectively, subject to their stated assumptions.

Proof. Expand the actual update $w_{t+1} = w_t - \eta(d_t + e_t)$. The error-free terms satisfy Equation (135). Bounding the remaining terms with $\|d_t\|_2 \leq D_{\mathrm{F}} \|G_t\|_2$ gives

$$F_N(w_{t+1}) \leq F_N(w_t) - \eta H_{\mathrm{F}} \|G_t\|_2^2 + \eta(1 + \beta\eta D_{\mathrm{F}}) \|G_t\|_2 \varepsilon_t + \frac{\beta\eta^2}{2} \varepsilon_t^2. \tag{277}$$

Young's inequality yields

$$(1 + \beta\eta D_{\mathrm{F}}) \|G_t\|_2 \varepsilon_t \leq \frac{H_{\mathrm{F}}}{2} \|G_t\|_2^2 + \frac{(1 + \beta\eta D_{\mathrm{F}})^2}{2H_{\mathrm{F}}} \varepsilon_t^2. \tag{278}$$

Using PL gives $\Delta_{t+1} \leq (1 - \mu\eta H_{\mathrm{F}}) \Delta_t + C_\eta \varepsilon_t^2$. Unrolling this recursion on the simultaneous event proves Equation (276). □

The task-loss gain survives in the contraction term, while the accumulated direction error controls the residual. If $\varepsilon_t \leq \bar{\varepsilon}$ uniformly, the residual is at most $C_\eta \bar{\varepsilon}^2 / (1 - q_{\mathrm{mb}})$. Vanishing error permits

convergence to the optimum under the same uniform conditions; fixed nonvanishing noise only gives a neighborhood guarantee. The stochastic result alone does not rank actual training curves or establish a lower error floor than another method.

B.6.2 High-probability comparison of stochastic orders

The deterministic distinction between a task-relative auxiliary field and a nonvanishing additive field persists under mini-batch sampling. For a method M, let d_t^{M} be its one-step full-data oracle direction and $\widehat{d}_t^{\mathsf{M}}$ its actual mini-batch direction. Suppose a simultaneous event $\mathcal{E}_{\mathsf{M}}$ satisfies

$$\|\widehat{d}_t^{\mathsf{M}} - d_t^{\mathsf{M}}\|_2 \le \varepsilon_{\mathsf{M},t}, \qquad 0 \le t < T. \tag{279}$$

For FORTE, this is the event produced by Lemma B.40 and the tower/union argument in Theorem B.41. Uniform and learned baselines have the analogous graph-sampling radius. An end-to-end learned gate must also concentrate its gate-Jacobian term in Equation (102).

Theorem B.43 (High-probability rate for a relative comparator)**.** *Suppose Assumption B.6 holds and, for every $t < T$, the full-data oracle direction satisfies*

$$G_t^\top d_t^{\mathsf{M}} \ge c_{\mathsf{M}} \|G_t\|_2^2, \qquad \|d_t^{\mathsf{M}}\|_2 \le D_{\mathsf{M}} \|G_t\|_2, \qquad c_{\mathsf{M}} > 0. \tag{280}$$

Choose

$$0 < \eta < \frac{2c_{\mathsf{M}}}{\beta D_{\mathsf{M}}^2}, \qquad a_{\mathsf{M}} = c_{\mathsf{M}} - \frac{\beta\eta D_{\mathsf{M}}^2}{2} > 0, \qquad b_{\mathsf{M}} = 1 + \beta\eta D_{\mathsf{M}}. \tag{281}$$

Then, on $\mathcal{E}_{\mathsf{M}}$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \|G_t\|_2^2 \le \frac{2\Delta_0}{\eta a_{\mathsf{M}} T} + \frac{1}{a_{\mathsf{M}} T}\left(\frac{b_{\mathsf{M}}^2}{a_{\mathsf{M}}} + \beta\eta\right) \sum_{t=0}^{T-1} \varepsilon_{\mathsf{M},t}^2. \tag{282}$$

Proof. On $\mathcal{E}_{\mathsf{M}}$, define

$$e_t^{\mathsf{M}} = \widehat{d}_t^{\mathsf{M}} - d_t^{\mathsf{M}}, \qquad \|e_t^{\mathsf{M}}\|_2 \le \varepsilon_{\mathsf{M},t}. \tag{283}$$

The actual update is $w_{t+1} = w_t - \eta(d_t^{\mathsf{M}} + e_t^{\mathsf{M}})$. Smoothness gives

$$F_N(w_{t+1}) \le F_N(w_t) - \eta G_t^\top (d_t^{\mathsf{M}} + e_t^{\mathsf{M}}) + \frac{\beta\eta^2}{2}\|d_t^{\mathsf{M}} + e_t^{\mathsf{M}}\|_2^2. \tag{284}$$

Use Equation (280) and $-G_t^\top e_t^{\mathsf{M}} \le \|G_t\|_2 \varepsilon_{\mathsf{M},t}$. Also,

$$\|d_t^{\mathsf{M}} + e_t^{\mathsf{M}}\|_2^2 \le \left(D_{\mathsf{M}}\|G_t\|_2 + \varepsilon_{\mathsf{M},t}\right)^2 \tag{285}$$

$$= D_{\mathsf{M}}^2\|G_t\|_2^2 + 2D_{\mathsf{M}}\|G_t\|_2\varepsilon_{\mathsf{M},t} + \varepsilon_{\mathsf{M},t}^2. \tag{286}$$

Substitution into Equation (284) yields

$$F_N(w_{t+1}) \le F_N(w_t) - \eta\left(c_{\mathsf{M}} - \frac{\beta\eta D_{\mathsf{M}}^2}{2}\right)\|G_t\|_2^2 \tag{287}$$

$$+ \eta(1 + \beta\eta D_{\mathsf{M}})\|G_t\|_2\varepsilon_{\mathsf{M},t} + \frac{\beta\eta^2}{2}\varepsilon_{\mathsf{M},t}^2 \tag{288}$$

$$= F_N(w_t) - \eta a_{\mathsf{M}}\|G_t\|_2^2 + \eta b_{\mathsf{M}}\|G_t\|_2\varepsilon_{\mathsf{M},t} + \frac{\beta\eta^2}{2}\varepsilon_{\mathsf{M},t}^2. \tag{289}$$

Young's inequality gives

$$b_{\mathsf{M}}\|G_t\|_2\varepsilon_{\mathsf{M},t} \le \frac{a_{\mathsf{M}}}{2}\|G_t\|_2^2 + \frac{b_{\mathsf{M}}^2}{2a_{\mathsf{M}}}\varepsilon_{\mathsf{M},t}^2. \tag{290}$$

Hence

$$F_N(w_{t+1}) \le F_N(w_t) - \frac{\eta a_{\mathsf{M}}}{2}\|G_t\|_2^2 + \frac{\eta}{2}\left(\frac{b_{\mathsf{M}}^2}{a_{\mathsf{M}}} + \beta\eta\right)\varepsilon_{\mathsf{M},t}^2. \tag{291}$$

Because $\mathcal{E}_{\mathsf{M}}$ is simultaneous, this inequality holds for all $t < T$ on the same sample path. Summing gives

$$\frac{\eta a_{\mathsf{M}}}{2} \sum_{t=0}^{T-1} \|G_t\|_2^2 \le \sum_{t=0}^{T-1} [F_N(w_t) - F_N(w_{t+1})] \tag{292}$$

$$+ \frac{\eta}{2} \left(\frac{b_{\mathsf{M}}^2}{a_{\mathsf{M}}} + \beta\eta \right) \sum_{t=0}^{T-1} \varepsilon_{\mathsf{M},t}^2 \tag{293}$$

$$= F_N(w_0) - F_N(w_T) \tag{294}$$

$$+ \frac{\eta}{2} \left(\frac{b_{\mathsf{M}}^2}{a_{\mathsf{M}}} + \beta\eta \right) \sum_{t=0}^{T-1} \varepsilon_{\mathsf{M},t}^2 \tag{295}$$

$$\le \Delta_0 + \frac{\eta}{2} \left(\frac{b_{\mathsf{M}}^2}{a_{\mathsf{M}}} + \beta\eta \right) \sum_{t=0}^{T-1} \varepsilon_{\mathsf{M},t}^2. \tag{296}$$

Division by $\eta a_{\mathsf{M}} T/2$ proves Equation (282). □

Theorem B.44 (High-probability floor for an additive comparator)**.** *Suppose Assumption B.6 holds and the full-data auxiliary field satisfies $\|V_t^{\mathsf{M}}\|_2 \le B_{\mathsf{M}}$. Suppose also that Equation (279) holds on $\mathcal{E}_{\mathsf{M}}$. Use $\eta = 1/(2\beta)$. Then, on $\mathcal{E}_{\mathsf{M}}$,*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|G_t\|_2^2 \le \frac{16\beta\Delta_0}{3T} + \frac{10}{9T} \sum_{t=0}^{T-1} (\lambda B_{\mathsf{M}} + \varepsilon_{\mathsf{M},t})^2. \tag{297}$$

Consequently,

$$\frac{1}{T} \sum_{t=0}^{T-1} \|G_t\|_2^2 \le \frac{16\beta\Delta_0}{3T} + \frac{20}{9}\lambda^2 B_{\mathsf{M}}^2 + \frac{20}{9T} \sum_{t=0}^{T-1} \varepsilon_{\mathsf{M},t}^2. \tag{298}$$

Increasing the mini-batch size removes the stochastic term but does not remove the nonvanishing auxiliary floor.

Proof. Write

$$\widehat{d}_t^{\mathsf{M}} = d_t^{\mathsf{M}} + (\widehat{d}_t^{\mathsf{M}} - d_t^{\mathsf{M}}) \tag{299}$$

$$= G_t + \lambda V_t^{\mathsf{M}} + e_t^{\mathsf{M}} \tag{300}$$

$$= G_t + \widetilde{e}_t, \tag{301}$$

where, on $\mathcal{E}_{\mathsf{M}}$,

$$\|\widetilde{e}_t\|_2 \le \lambda \|V_t^{\mathsf{M}}\|_2 + \|e_t^{\mathsf{M}}\|_2 \le \lambda B_{\mathsf{M}} + \varepsilon_{\mathsf{M},t}. \tag{302}$$

Applying smoothness to $w_{t+1} = w_t - \eta(G_t + \widetilde{e}_t)$ gives the exact identity

$$F_N(w_{t+1}) \le F_N(w_t) - \eta \left(1 - \frac{\beta\eta}{2}\right) \|G_t\|_2^2 - \eta(1 - \beta\eta) G_t^\top \widetilde{e}_t + \frac{\beta\eta^2}{2} \|\widetilde{e}_t\|_2^2. \tag{303}$$

The Young-inequality calculation in Equations (97)–(98), now with e_t replaced by $\widetilde{e}_t$, yields at $\eta = 1/(2\beta)$

$$F_N(w_{t+1}) \le F_N(w_t) - \frac{3\eta}{8} \|G_t\|_2^2 + \frac{5\eta}{12} \|\widetilde{e}_t\|_2^2. \tag{304}$$

Sum over t, telescope F_N, use Equation (302), and divide by $3\eta T/8$:

$$\frac{1}{T} \sum_{t=0}^{T-1} \|G_t\|_2^2 \le \frac{8\Delta_0}{3\eta T} + \frac{10}{9T} \sum_{t=0}^{T-1} (\lambda B_{\mathsf{M}} + \varepsilon_{\mathsf{M},t})^2 \tag{305}$$

$$= \frac{16\beta\Delta_0}{3T} + \frac{10}{9T} \sum_{t=0}^{T-1} (\lambda B_{\mathsf{M}} + \varepsilon_{\mathsf{M},t})^2. \tag{306}$$

This proves Equation (297). Finally, $(x+y)^2 \le 2x^2 + 2y^2$ gives Equation (298). □

Corollary B.45 (Stochastic order comparison)**.** *Assume the direction radii obey*

$$\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon_{\mathsf{M},t}^2 = \widetilde{\mathcal{O}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{b_t} + \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k}\frac{(\alpha_{k,t}^{\mathrm{EMA}})^2}{n_{k,t}}\right). \tag{307}$$

Then FORTE and any comparator satisfying relative oracle geometry have

$$\mathcal{O}(T^{-1}) + \widetilde{\mathcal{O}}\left(\frac{1}{T}\sum_{t}\frac{1}{b_t} + \frac{1}{T}\sum_{t}\sum_{k}\frac{(\alpha_{k,t}^{\mathrm{EMA}})^2}{n_{k,t}}\right), \tag{308}$$

whereas an additive comparator has the same two decaying terms plus $\mathcal{O}(\lambda^2 B_{\mathsf{M}}^2)$. If $b_t, n_{k,t} = \widetilde{\Theta}(\sqrt{T})$, the decaying part is $\widetilde{\mathcal{O}}(T^{-1/2})$. With full batches, it is $\mathcal{O}(T^{-1})$.

Proof. Substitute Equation (307) into Theorem B.43 for a relative method and into Theorem B.44 for an additive method. The claimed special cases follow by evaluating the displayed averages. □

Corollary B.46 (Explicit mini-batch order)**.** *Suppose the Bernstein and GNN-Jacobian quantities are uniformly bounded and each available teacher contributes at least $b_{\min}$ observations whenever it is updated. Then, up to logarithmic factors in $T(K+1)/\delta_{\mathrm{mb}}$,*

$$\varepsilon_t^2 = \widetilde{\mathcal{O}}\left(\frac{1}{b_t} + \sum_{k=1}^{K}\frac{(\alpha_{k,t}^{\mathrm{EMA}})^2}{n_{k,t}}\right). \tag{309}$$

Consequently, with a constant stable step size,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_N(w_t)\|_2^2 = \mathcal{O}(T^{-1}) + \widetilde{\mathcal{O}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{b_t} + \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k}\frac{(\alpha_{k,t}^{\mathrm{EMA}})^2}{n_{k,t}}\right). \tag{310}$$

For fixed mini-batch size this is convergence to a variance-controlled neighborhood. If $b_t, n_{k,t} = \widetilde{\Theta}(\sqrt{T})$, the bound is $\widetilde{\mathcal{O}}(T^{-1/2})$; if they are $\widetilde{\Theta}(T)$, it is $\widetilde{\mathcal{O}}(T^{-1})$.

Proof. Square Equation (252), use $(x+y)^2 \le 2x^2 + 2y^2$, and substitute Equations (249)–(250). The leading variance terms scale as b_t^{-1} and $(\alpha_{k,t}^{\mathrm{EMA}})^2 n_{k,t}^{-1}$; the Bernstein linear-tail terms are of order b_t^{-2} and $n_{k,t}^{-2}$ up to squared logarithms. Substitute this result into Theorem B.41. □

B.7 Why the stochastic assumptions hold for bounded finite-depth GNNs

Proposition B.47 (A concrete bounded-increment GNN regime)**.** *Consider the normalized finite-depth message-passing GNN in Equation (225). Suppose, on the analyzed finite training horizon,*

$$\|\widetilde{A}_i\|_{\mathrm{op}} \le 1, \quad \|X_i\|_F \le B_X, \quad \|P_i\|_{\mathrm{op}} \le B_P, \quad \|W_\ell\|_{\mathrm{op}} \le M_\ell, \quad \sup_z |\sigma_\ell'(z)| \le L_\ell. \tag{311}$$

Suppose the task-head Jacobian is bounded by M_q, the task-loss derivative at the head output is bounded by B_ℓ, and teacher-k's private head and loss derivative are bounded by $M_{p,k}$ and $B_{D,k}$. Then all per-graph oracle parameter directions and all teacher gradient-norm observations are bounded. Hence the Bernstein moment conditions in Definition B.39 hold as a direct special case, with no Gaussian assumption.

Proof. First bound the forward features. Let $B_0 = B_X$ and recursively define

$$B_\ell = L_\ell M_\ell B_{\ell-1}. \tag{312}$$

Because $\|\widetilde{A}_i\|_{\mathrm{op}} \le 1$ and the activation is L_ℓ-Lipschitz,

$$\|H_i^{(\ell)}\|_F \le L_\ell \|\widetilde{A}_i H_i^{(\ell-1)} W_\ell\|_F \le L_\ell M_\ell \|H_i^{(\ell-1)}\|_F \le B_\ell. \tag{313}$$

Next bound the Jacobian block associated with W_s. A perturbation ΔW_s first produces

$$\|\Delta H_i^{(s)}\|_F \le L_s \|H_i^{(s-1)}\|_F \|\Delta W_s\|_F \le L_s B_{s-1} \|\Delta W_s\|_F. \tag{314}$$

Each subsequent layer multiplies this perturbation by at most $L_\ell M_\ell$. Pooling contributes B_P. Therefore

$$\|D_{W_s}\boldsymbol{h}_i\|_{\mathrm{op}} \le B_P L_s B_{s-1} \prod_{\ell=s+1}^{L} L_\ell M_\ell =: C_s. \tag{315}$$

For all GNN blocks together,

$$\|\boldsymbol{J}_i\|_{\mathrm{op}} \le \left(\sum_{s=1}^{L} C_s^2\right)^{1/2} =: B_J. \tag{316}$$

The task representation gradient satisfies

$$\|\boldsymbol{g}_i\|_2 \le \|D_h q_\phi(\boldsymbol{h}_i)\|_{\mathrm{op}} \|\nabla_q \ell_{\mathrm{sup}}\|_2 \le M_q B_\ell =: B_g. \tag{317}$$

Likewise,

$$\|\boldsymbol{u}_{k,i}\|_2 \le M_{p,k} B_{D,k} =: B_{u,k}. \tag{318}$$

The unchanged calibration and evidence gate give $\|\boldsymbol{a}_i\|_2 \le \|\boldsymbol{g}_i\|_2$. Hence the shared-GNN part of the per-graph routed direction obeys

$$\|\boldsymbol{J}_i^\top(\boldsymbol{g}_i + \lambda \boldsymbol{a}_i)\|_2 \le B_J(1+\lambda)B_g. \tag{319}$$

Adding the bounded task-head gradient gives a finite bound B_d on the complete deployed-parameter direction. Centering at its finite-population mean multiplies the bound by at most two. Therefore, for every integer $m \ge 2$,

$$\|\xi_i\|_2^m \le (2B_d)^{m-2}\|\xi_i\|_2^2. \tag{320}$$

Summation of conditional expectations yields the Bernstein moment condition with scale $\Gamma_{d,t} = 2B_d$ and the actual conditional second-moment sum as variance proxy. The same argument applies to the scalar observations $\|\boldsymbol{u}_{k,i}\|_2 \le B_{u,k}$. Thus bounded normalized GNNs are a concrete special case of the concentration assumptions. □

Interpretation. The corrected proof has two separate probability layers. Width or random initialization controls the full-data and realized mini-batch GNN tangent metrics through Propositions B.31 and B.33. Conditional on that event, graph subsampling and the current-batch EMA are controlled directly in the actual parameter space by Theorem B.41. No convergence statement treats the representation as an independently updated optimization variable.

C Additional Experimental Material

C.1 Implementation Details

We use GIN, MolCLR, and Mole-BERT as the 2D molecular graph students and SMI, IEM, and EDG as the three heterogeneous teachers. The classification configuration uses five GIN message-passing layers with 300-dimensional hidden features and the final-layer representation. The three backbones load EdgePred, MolCLR-GIN, and Mole-BERT checkpoints, respectively. The pretrained teacher encoders remain frozen, and the 256-dimensional teacher representations are cached together with their availability masks. For molecule i, the student encoder produces a representation $\boldsymbol{h}_i$. For each teacher k, a teacher-specific projection head p_{ψ_k} maps $\boldsymbol{h}_i$ into the corresponding teacher representation space of $\boldsymbol{z}_{k,i}$. Each projection head is implemented as $d \longrightarrow 512 \longrightarrow 256$, consisting of a linear layer, LayerNorm, ReLU, dropout with $p = 0.1$, and a second linear layer. Here, d denotes the dimensionality of the student representation (e.g., $d = 300$ for GIN), while all three cached teacher representations are 256-dimensional. The projection heads are optimized jointly with the student during distillation, whereas the cached teacher representations remain fixed. For FORTE routing, we use $\rho = 0.99$, $c_{\max} = 10$, $\gamma = 0$, $\epsilon_c = \epsilon_\omega = 10^{-8}$. The routing temperature τ is selected from $\{0.02, 0.05, 0.10, 0.20\}$, and the distillation strength λ is selected from $\{0.10, 0.25, 0.50, 1.00\}$. This gives a total of $4 \times 4 = 16$ hyperparameter configurations. Hyperparameters are selected exclusively according to validation performance. The final configuration uses $\tau = 0.10$, $\lambda = 1.00$. All remaining FORTE parameters are kept fixed. For the main classification experiments, FORTE and the matched comparison methods use the same scaffold split, student backbone, training schedule, and random seed. Models are trained for 100 epochs with a batch size of 32 and a learning rate of 10^{-3}. For the regression experiments reported in Table 5, all methods are trained for 100 epochs. In both settings, model and checkpoint selection are performed exclusively on the validation set, and the test set is used only for final evaluation. We repeat each experiment with random seeds 0, 1, and 2 and report the mean and standard deviation across the three runs. The comparison methods follow the unified reproduction protocol described in Appendix C.3.

C.2 Datasets of Downstream Tasks

The ESOL, Lipo, Malaria, and CEP datasets are from GraphMVP (Liu et al., 2022). The other datasets are from MoleculeNet (Wu et al., 2018). Table 3 summarizes the eight classification and four regression benchmarks.

Table 3: Overview of the twelve downstream molecular property benchmarks. Molecule and task counts follow the dataset releases listed in the official IEM repository (Xiang et al., 2024).

Type	Dataset	Molecules	Tasks	Prediction target	Metric
Classification	Tox21	7,831	12	Toxicity in 12 biological pathways	ROC-AUC ↑
	ToxCast	8,576	617	High-throughput toxicity screening	ROC-AUC ↑
	SIDER	1,427	27	Adverse drug reactions	ROC-AUC ↑
	ClinTox	1,478	2	Clinical toxicity and FDA approval	ROC-AUC ↑
	MUV	93,087	17	Activity in imbalanced assays	ROC-AUC ↑
	HIV	41,127	1	Inhibition of HIV replication	ROC-AUC ↑
	BBBP	2,039	1	Blood–brain barrier penetration	ROC-AUC ↑
	BACE	1,513	1	BACE-1 inhibition	ROC-AUC ↑
Regression	ESOL	1,128	1	Aqueous solubility	RMSE ↓
	Lipo	4,200	1	Lipophilicity (log D, pH 7.4)	RMSE ↓
	Malaria	9,999	1	Antimalarial activity	RMSE ↓
	CEP	29,978	1	Clean Energy Project properties	RMSE ↓

Source: IEM official dataset listing.

Further descriptions of the datasets are provided below:

- **BBBP (Blood-Brain Barrier Penetration)** records the relationship between drugs and barrier permeability, which is important for assessing whether drugs are blocked by the blood-brain barrier.

- **Tox21 (Toxicology in the 21st Century)** contains qualitative toxicity measurements of compounds on 12 different targets, including nuclear receptors and stress response pathways.
- **ClinTox** is a compound toxicity dataset that includes FDA-approved drugs and drugs that failed clinical trials for toxicity reasons.
- **BACE (Beta-secretase)** provides qualitative binary results for compounds targeting human β-secretase 1 (BACE-1).
- **SIDER (Side Effect Resource)** is a dataset of marketed drugs and adverse drug reactions with 27 system organ classes.
- **ToxCast** extends Tox21 with toxicology data from a large compound library based on in vitro high-throughput screening.
- **HIV** provides compounds from the Drug Therapeutics Program AIDS Antiviral Screen that inhibit HIV replication.
- **ESOL** records aqueous solubility and is widely used to validate models that estimate solubility from molecular structure.
- **Lipo (Lipophilicity)** provides experimental octanol/water distribution coefficients ($\log D$ at pH 7.4), which affect membrane permeability and solubility.
- **Malaria** measures drug efficacy against the parasite that causes malaria.
- **CEP** is a subset of the Harvard Clean Energy Project.

Figure 6 illustrates the inputs associated with the student and three teacher modalities. The 2D graph is available to the student at training and inference time. IEM supplies image-based conformer views during training. EDG supplies features predicted from structural images under electron-density supervision during pre-training, rather than requiring an electron-density map for each downstream molecule. The language teacher receives the corresponding SMILES sequence. Only the 2D graph is required after distillation.

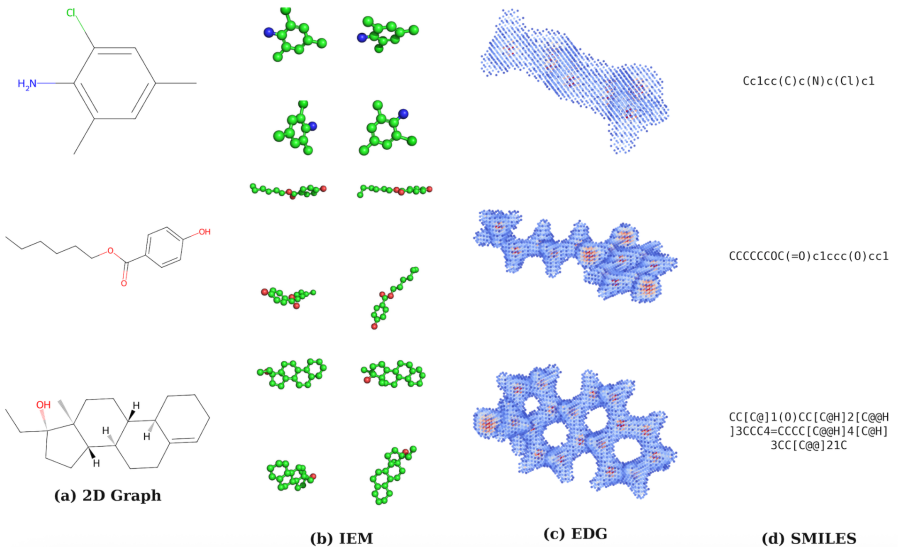

Figure 6: The same molecule is described through graph connectivity, conformer geometry, electron density, and a SMILES sequence.

C.3 Competitors

We compare FORTE with methods that differ in how they represent the molecule, transfer teacher knowledge, or combine multiple training signals. The methods in Tables 1 and 5 are grouped according to their underlying mechanism. An unprefixed backbone is a student-only model, whereas a "+" indicates that the corresponding method is applied to that student.

1) Student-only molecular graph models.

- **GIN** (Xu et al., 2019) is a graph isomorphism network that updates each atom representation by aggregating the representations of its neighbors. Its expressive message-passing scheme provides a supervised 2D graph baseline.
- **MolCLR** (Wang et al., 2022) pre-trains a molecular graph encoder with contrastive learning over augmented views of the same molecule. It tests whether teacher supervision remains useful when the student already has a representation learned from unlabeled molecular graphs.
- **Mole-BERT** (Xia et al., 2023) uses context-aware atom tokenization and masked graph modeling to learn molecular representations. It provides a second pre-trained graph student with a different pre-training objective from MolCLR.

2) Molecular knowledge distillation methods.

- **CCMD** (Zhang et al., 2022) transfers knowledge between complementary molecular modalities for property prediction. Its central idea is to use cross-modal supervision to improve a molecular representation that can be used by the downstream predictor.
- **MolKD** (Zeng et al., 2023) transfers molecular knowledge from a richer teacher representation to a property-prediction student through distillation. It represents the class of methods that use a pre-trained teacher to enrich a deployable molecular predictor.
- **IEM** (Xiang et al., 2024) uses image-based molecular representations together with graph representations to capture molecular properties and geometric information. Its task enhancer transfers visual and geometric knowledge to the graph student.
- **D&D** (Cho et al., 2025) uses a 3D denoising model as the teacher for a 2D molecular representation. The student can therefore receive geometric knowledge during training while retaining a 2D input path.

3) Teacher prediction fusion. These methods first combine the predictions of the available teachers and then train the student against the resulting target. They test whether a fused teacher prediction is sufficient for heterogeneous distillation.

- **Uniform-Fused** averages the available teacher predictions with equal weights. Every teacher contributes to the fused target regardless of its compatibility with the downstream task.
- **MoE-Fused** forms a mixture-of-experts target using learned, sample-dependent gate weights. It allows the mixture to vary across molecules without using the downstream task gradient to determine teacher eligibility.

4) Teacher-wise distillation loss weighting. These methods keep a separate distillation loss for each teacher and combine the losses with scalar or sample-dependent weights. Unlike prediction fusion, they preserve the individual teacher objectives before aggregation.

- **Uniform-KD** assigns equal weights to the available teacher losses. It is the standard multi-teacher distillation control and treats all available supervision as equally useful.
- **Learned-Gate** combines teacher losses using learned, sample-dependent gate weights. Its gate adapts the mixture from the available teacher information without explicitly measuring alignment with the task gradient.
- **Confidence-Gate** weights each teacher according to its predictive confidence. For classification, confidence is derived from the normalized entropy of the teacher predictions, so the method selects confident teachers without directly testing whether their updates support task descent.

5) Gradient balancing and conflict-aware optimization. These methods operate on gradients from multiple objectives rather than only on their scalar losses. They provide controls for separating the effect of gradient magnitude balancing from FORTE's task-prioritized routing.

- **GradNorm** (Chen et al., 2018) adapts objective weights according to gradient magnitudes and relative training rates. It balances the optimization influence of the teacher objectives without making teacher participation conditional on task compatibility.
- **PCGrad** (Yu et al., 2020) projects a gradient away from another objective when the two gradients conflict. It resolves pairwise conflicts among objectives, while its decision rule is based on objective gradients rather than a routed teacher-to-task contribution.
- **CAGrad** (Liu et al., 2021) constructs a conflict-averse aggregate gradient that balances improvement across objectives. It treats the objectives symmetrically, whereas the downstream property loss is the primary objective in our distillation setting.

Reproduction Protocol for Competitors. All comparison methods are evaluated under a unified reproduction protocol. Each model is trained for 100 epochs with three random seeds (0, 1 and 2). Within each dataset–backbone setting, all methods share the same data splits and student initialization. Batch size, learning rate and training schedule are also held fixed. Published comparison methods are reproduced using the authors' released code and the descriptions in their respective papers, with shared experimental settings standardized to this protocol. Method-specific mechanisms are retained as described above. Hyperparameters and checkpoints are selected exclusively using the validation set. The test set is reserved for final evaluation, and results are reported as mean and standard deviation across the three runs. Further training settings are provided in Appendix C.1.

C.4 Comparison with More State-of-the-Art Methods

Table 4 provides a broader comparison with additional state-of-the-art methods in the GIN setting. Among the methods shown, FORTE achieves the highest average ROC-AUC across the 8 classification datasets.

Table 4: ROC-AUC (%) performance on 8 molecular property classification datasets in the GIN setting. We report the mean and standard deviation over 3 random seeds under scaffold splitting. The Average column reports the arithmetic mean of the 8 dataset means. Best and second-best mean values are bold and underlined, respectively.

Method	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BBBP	BACE	Average
#Molecules	7831	8576	1427	1478	93087	41127	2039	1513	–
#Task	12	617	27	2	17	1	1	1	–
InfoGraph (Sun et al., 2020)	73.3 ± 0.6	61.8 ± 0.4	58.7 ± 0.6	75.4 ± 4.3	74.4 ± 1.8	74.2 ± 0.9	68.7 ± 0.6	74.3 ± 2.6	70.1
GPT-GNN (Hu et al., 2020b)	74.9 ± 0.3	62.5 ± 0.4	58.1 ± 0.3	58.3 ± 5.2	75.9 ± 2.3	65.2 ± 2.1	64.5 ± 1.4	77.9 ± 3.2	68.5
ContextPred (Hu et al., 2020a)	73.6 ± 0.3	62.6 ± 0.6	59.7 ± 1.8	74.0 ± 3.4	72.5 ± 1.5	75.6 ± 1.0	70.6 ± 1.5	78.8 ± 1.2	70.9
G-Contextual (Rong et al., 2020)	75.0 ± 0.6	62.8 ± 0.7	58.7 ± 1.0	60.6 ± 5.2	72.1 ± 0.7	76.3 ± 1.5	69.9 ± 2.1	79.3 ± 1.1	69.3
GraphLoG (Xu et al., 2021)	75.0 ± 0.6	63.4 ± 0.6	59.6 ± 1.9	75.7 ± 2.4	75.5 ± 1.6	76.1 ± 0.8	68.7 ± 1.6	78.6 ± 1.0	71.6
FG-BERT (Li et al., 2023)	78.4 ± 0.8	66.3 ± 0.8	64.0 ± 0.7	83.2 ± 1.6	75.3 ± 2.4	77.4 ± 1.0	70.2 ± 0.9	**84.5 ± 1.5**	74.9
MoleculeSDE (Liu et al., 2023)	76.8 ± 0.3	65.0 ± 0.2	60.8 ± 0.3	87.0 ± 0.5	80.9 ± 0.3	78.8 ± 0.9	71.8 ± 0.7	79.5 ± 2.1	75.1
MoleMCL (Zhang et al., 2024)	76.2 ± 0.6	63.7 ± 0.2	59.0 ± 0.8	80.2 ± 1.6	78.2 ± 1.0	**79.6 ± 0.5**	67.8 ± 3.5	80.3 ± 0.8	73.1
MOLEBLEND (Yu et al., 2024)	77.8 ± 0.8	66.1 ± 0.0	64.9 ± 0.3	**87.6 ± 0.7**	77.2 ± 2.3	79.0 ± 0.8	**73.0 ± 0.8**	83.7 ± 1.4	76.2
GIN	76.4 ± 0.3	63.8 ± 0.2	60.5 ± 0.4	66.3 ± 1.6	74.2 ± 0.7	77.5 ± 0.8	66.2 ± 3.9	79.9 ± 1.5	70.6
+Du-GCS (Du et al., 2018)	76.7 ± 1.1	64.7 ± 1.3	60.1 ± 0.5	63.7 ± 3.4	76.1 ± 2.9	76.2 ± 1.8	65.7 ± 0.6	79.8 ± 1.4	70.4
+MoE-Fused	76.8 ± 1.0	64.0 ± 1.3	60.5 ± 0.3	63.2 ± 0.7	77.7 ± 1.9	78.2 ± 1.7	68.0 ± 2.5	80.1 ± 0.8	71.1
+Confidence-Gate	76.6 ± 0.6	63.6 ± 0.2	60.9 ± 0.7	61.4 ± 1.1	76.1 ± 2.5	77.4 ± 1.4	66.0 ± 1.7	79.9 ± 0.8	70.2
+CAGrad (Liu et al., 2021)	75.8 ± 0.8	64.0 ± 0.6	60.7 ± 0.7	66.5 ± 4.8	70.8 ± 1.9	77.8 ± 1.5	70.7 ± 1.9	79.3 ± 0.3	70.7
+D&D (Cho et al., 2025)	76.3 ± 0.2	64.4 ± 0.1	60.0 ± 0.2	65.6 ± 1.6	**81.0 ± 1.1**	77.4 ± 1.2	64.6 ± 2.9	80.7 ± 2.0	71.3
+FORTE	**79.4 ± 0.5**	**67.6 ± 0.2**	**66.1 ± 0.4**	86.1 ± 3.3	78.7 ± 1.0	79.3 ± 3.1	**73.0 ± 5.6**	81.6 ± 1.2	**76.5**

C.5 Teacher Sources and Feature Extraction

Pre-trained weight sources. We use existing pre-trained weights for all three heterogeneous teachers and do not pre-train their encoders as part of FORTE. The conformational-image teacher IEM is initialized from the author-released `IEM.pth` checkpoint of Xiang et al. (Xiang et al., 2024). The electron-density-aware teacher EDG uses the author-released `ED-aware-Teacher.pth` checkpoint, which contains the `ED_teacher` encoder and `EDPredictor` weights (Xiang et al., 2025). The SMILES teacher SMI is initialized from the public `DeepChem/ChemBERTa-77M-MTR` checkpoint, following the ChemBERTa family of SMILES encoders (Chithrananda et al., 2020). These source checkpoints remain frozen during feature extraction.

How the source weights were pre-trained. The three checkpoints encode different forms of molecular knowledge. The ChemBERTa weights were learned from SMILES sequences using large-scale molecular language pre-training. The released ChemBERTa resources use a corpus of 77 million PubChem SMILES strings (Chithrananda et al., 2020). IEM was pre-trained on two-dimensional and multi-view three-dimensional molecular images (Xiang et al., 2024). Its objectives combine cross-view contrastive learning with prediction of molecular properties and geometric quantities. These geometric quantities include atom distances and bond distances. EDG was pre-trained on two million molecules to transfer electron-density information into a structural-image encoder (Xiang et al., 2025). Its `EDPredictor` maps pooled structural-image features toward electron-density representations extracted from DFT-derived supervision. These original pre-training procedures produce the weights that we load. They are not repeated in our downstream experiments.

SMI feature extraction. Each molecule is represented by its SMILES string and tokenized with the released ChemBERTa tokenizer. The frozen encoder produces a sequence of hidden states. We use the final hidden state of the leading token as the raw SMI feature, rather than using property predictions or task logits. No image or molecular conformation is required for this teacher.

IEM feature extraction. A molecule's SMILES string is used to construct a three-dimensional conformation. The conformation is rendered into multiple molecular images from different viewpoints. The frozen ResNet18-based `image3d_teacher` from the released IEM checkpoint encodes each view. Images are resized to 224×224 pixels and processed with fixed ImageNet normalization. For the available views $\mathcal{V}_i$ of molecule i, the raw teacher feature is

$$\boldsymbol{x}_{\mathrm{IEM},i}^{\mathrm{raw}} = \frac{1}{|\mathcal{V}_i|} \sum_{v \in \mathcal{V}_i} E_{\mathrm{IEM}}(x_{iv}), \tag{321}$$

where x_{iv} is a rendered view and E_{IEM} is frozen. The downstream cache uses the pooled feature from the 3D image branch. It does not use the auxiliary prediction heads from IEM pre-training.

EDG feature extraction. EDG also starts from multi-view structural images rendered from a molecular conformation. The released checkpoint contains the frozen image encoder `ED_teacher` and its `EDPredictor`. Their original pre-training objective matches the predicted representation to a DFT-derived electron-density target:

$$\mathcal{L}_{\mathrm{EDG}} = \mathrm{MSE}\left(P_{\mathrm{ED}}\left(\frac{1}{|\mathcal{V}_i|} \sum_{v \in \mathcal{V}_i} E_{\mathrm{ED}}(x_{iv}) \right), z_i^{\text{DFT-ED}} \right). \tag{322}$$

Here E_{ED} denotes the structural-image encoder, P_{ED} the density-feature predictor, and $z_i^{\text{DFT-ED}}$ the pre-training target representation. Electron density is therefore the supervision used to create the released weights. It is not a raw input required by FORTE. During downstream feature extraction, we average the structural-image features and apply the frozen `EDPredictor`. This produces a 768-dimensional electron-density-aware feature. No new DFT electron-density calculation is performed for a downstream molecule.

C.6 Regression Results

We further evaluate FORTE on four regression benchmarks. Table 5 reports the corresponding RMSE results. FORTE obtains the lowest RMSE in 11 of 12 backbone–dataset settings. The maximum relative reduction over the corresponding 2D student is 9.95%. The only exception is CEP with Mole-BERT, where IEM gives a lower RMSE. These results show that the benefit of teacher routing extends beyond classification while preserving the 2D inference path.

Downstream optimization on regression tasks. To examine whether the optimization behavior extends to regression, we compare the task-loss trajectories of FORTE, Uniform KD, GradNorm, PCGrad, and CAGrad on ESOL, Lipophilicity, Malaria, and CEP in Figure 7. We observe a rapid initial decrease for all methods, followed by more gradual improvement over the remaining training epochs. FORTE generally follows a lower trajectory during the middle and late stages and ends with the lowest plotted task loss on all four datasets. The separation is more visible on ESOL and Lipophilicity, while the curves remain closer on Malaria and CEP. These trajectories support

Table 5: Regression RMSE on four molecular benchmarks with GIN, MolCLR, and Mole-BERT students.

Type	Method	ESOL		Lipophilicity		Malaria		CEP	
Student	**GIN**	1.38 ±0.04		0.80 ±0.02		1.11 ±0.01		1.37 ±0.04	
Fusion	+Uniform-Fused	1.34 ±0.02	0.04	0.77 ±0.00	0.04	1.10 ±0.01	0.01	1.37 ±0.02	0.00
	+MoE-Fused	1.36 ±0.02	0.01	0.78 ±0.01	0.03	1.10 ±0.01	0.00	1.37 ±0.05	0.00
Loss weighting	+Uniform-KD	1.37 ±0.03	0.01	0.77 ±0.00	0.03	1.10 ±0.01	0.01	1.37 ±0.03	0.00
	+Learned-Gate	1.35 ±0.03	0.03	0.78 ±0.01	0.03	1.11 ±0.02	0.00	1.37 ±0.03	0.00
	+Confidence-Gate	1.35 ±0.04	0.02	0.77 ±0.01	0.03	1.10 ±0.01	0.00	1.38 ±0.03	-0.01
Gradient optimization	+GradNorm	1.35 ±0.01	0.03	0.77 ±0.00	0.03	1.10 ±0.01	0.01	1.38 ±0.03	-0.01
	+PCGrad	1.35 ±0.03	0.03	0.77 ±0.01	0.03	1.09 ±0.01	0.01	1.38 ±0.03	-0.01
	+CAGrad	1.32 ±0.02	0.06	0.77 ±0.00	0.03	1.09 ±0.00	0.01	1.39 ±0.01	-0.02
Molecular distillation	+CCMD	1.45 ±0.03	-0.07	0.94 ±0.04	-0.14	1.15 ±0.03	-0.04	1.81 ±0.00	-0.44
	+MolKD	1.40 ±0.06	-0.03	0.80 ±0.02	0.00	1.12 ±0.02	-0.01	1.42 ±0.01	-0.04
	+IEM	1.36 ±0.04	0.02	0.78 ±0.01	0.02	1.12 ±0.01	-0.01	1.37 ±0.01	0.00
	+D&D	1.44 ±0.06	-0.06	0.85 ±0.03	-0.05	1.12 ±0.01	-0.01	1.54 ±0.01	-0.17
Ours	**FORTE**	**1.30** ±0.01	0.08	**0.73** ±0.01	0.07	**1.05** ±0.00	0.06	**1.29** ±0.02	0.08
Student	**MolCLR**	1.48 ±0.08		0.80 ±0.02		1.11 ±0.01		1.39 ±0.03	
Fusion	+Uniform-Fused	1.37 ±0.02	0.11	0.78 ±0.02	0.02	1.10 ±0.01	0.01	1.41 ±0.06	-0.01
	+MoE-Fused	1.38 ±0.06	0.10	0.79 ±0.01	0.01	1.10 ±0.01	0.01	1.40 ±0.04	-0.01
Loss weighting	+Uniform-KD	1.41 ±0.06	0.07	0.79 ±0.01	0.02	1.10 ±0.01	0.01	1.43 ±0.04	-0.03
	+Learned-Gate	1.41 ±0.04	0.07	0.76 ±0.02	0.04	1.10 ±0.00	0.01	1.41 ±0.05	-0.02
	+Confidence-Gate	1.35 ±0.04	0.13	0.78 ±0.01	0.02	1.09 ±0.00	0.01	1.38 ±0.03	0.01
Gradient optimization	+GradNorm	1.38 ±0.07	0.10	0.77 ±0.01	0.03	1.11 ±0.02	0.00	1.41 ±0.04	-0.01
	+PCGrad	1.38 ±0.04	0.10	0.78 ±0.01	0.02	1.09 ±0.01	0.01	1.39 ±0.01	0.01
	+CAGrad	1.41 ±0.10	0.07	0.79 ±0.01	0.01	1.09 ±0.00	0.01	1.43 ±0.02	-0.03
Molecular distillation	+CCMD	1.41 ±0.04	0.07	0.84 ±0.03	-0.04	1.17 ±0.00	-0.06	1.63 ±0.02	-0.23
	+MolKD	1.42 ±0.07	0.06	0.79 ±0.01	0.02	1.11 ±0.01	0.00	1.42 ±0.05	-0.03
	+IEM	1.51 ±0.15	-0.04	0.81 ±0.02	-0.01	1.11 ±0.01	-0.01	1.43 ±0.01	-0.03
	+D&D	1.46 ±0.10	0.02	0.84 ±0.04	-0.03	1.11 ±0.02	0.00	1.53 ±0.02	-0.14
Ours	**FORTE**	**1.33** ±0.01	0.15	**0.72** ±0.01	0.08	**1.04** ±0.00	0.06	**1.30** ±0.01	0.09
Student	**Mole-BERT**	1.09 ±0.01		0.72 ±0.01		1.13 ±0.00		1.35 ±0.01	
Fusion	+Uniform-Fused	1.10 ±0.06	-0.01	0.71 ±0.01	0.01	1.10 ±0.01	0.02	1.36 ±0.01	-0.02
	+MoE-Fused	1.09 ±0.05	-0.01	0.71 ±0.00	0.01	1.11 ±0.03	0.02	1.35 ±0.01	-0.01
Loss weighting	+Uniform-KD	1.11 ±0.06	-0.02	0.71 ±0.01	0.01	1.10 ±0.01	0.02	1.36 ±0.01	-0.02
	+Learned-Gate	1.09 ±0.05	-0.01	0.71 ±0.01	0.01	1.10 ±0.01	0.02	1.35 ±0.01	-0.01
	+Confidence-Gate	1.06 ±0.01	0.03	0.71 ±0.01	0.01	1.10 ±0.00	0.03	1.35 ±0.01	-0.01
Gradient optimization	+GradNorm	1.10 ±0.05	-0.01	0.71 ±0.01	0.01	1.10 ±0.01	0.03	1.36 ±0.01	-0.01
	+PCGrad	1.09 ±0.06	-0.01	0.71 ±0.01	0.01	1.11 ±0.01	0.02	1.36 ±0.01	-0.02
	+CAGrad	1.09 ±0.06	-0.01	0.71 ±0.01	0.01	1.11 ±0.01	0.02	1.38 ±0.02	-0.03
Molecular distillation	+CCMD	1.44 ±0.01	-0.36	0.93 ±0.03	-0.21	1.15 ±0.01	-0.02	1.84 ±0.02	-0.49
	+MolKD	1.12 ±0.03	-0.03	0.72 ±0.01	0.00	1.12 ±0.01	0.01	1.40 ±0.00	-0.05
	+IEM	1.09 ±0.01	-0.01	0.73 ±0.01	-0.01	1.15 ±0.02	-0.02	**1.34** ±0.01	0.01
	+D&D	1.11 ±0.03	-0.02	0.73 ±0.00	-0.01	1.14 ±0.02	-0.01	1.60 ±0.01	-0.25
Ours	**FORTE**	**1.04** ±0.02	0.05	**0.69** ±0.00	0.04	**1.05** ±0.00	0.07	1.38 ±0.01	-0.03

improved late-stage task optimization across the four regression tasks, complementing the test RMSE comparison in Table 5 rather than replacing it.

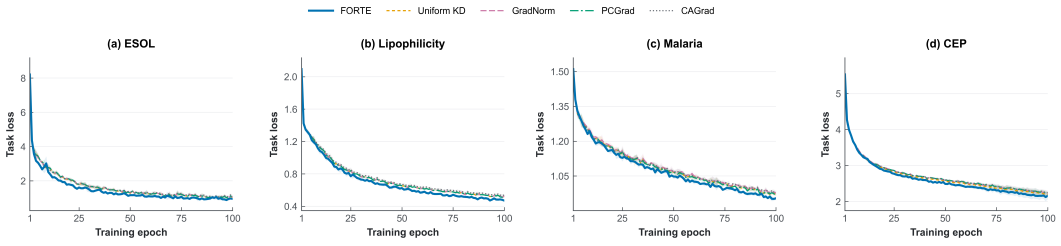

Figure 7: (a) ESOL, (b) Lipophilicity, (c) Malaria, and (d) CEP: task-loss trajectories of FORTE and four distillation baselines over 100 training epochs. FORTE ends with the lowest plotted task loss on each dataset, with clearer separation on ESOL and Lipophilicity than on Malaria and CEP.

C.7 Additional Task-Loss Trajectories

Figure 8 completes the task-loss comparison in Section 3.1 with results on MUV, Tox21, SIDER, and ToxCast. These plots show supervised training loss over the first 100 epochs.

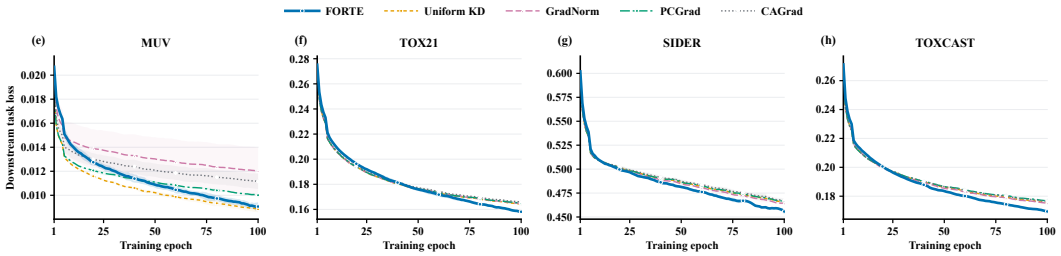

Figure 8: Supervised training loss on (e) MUV, (f) Tox21, (g) SIDER, and (h) ToxCast over 100 epochs, shown as three-seed mean ± SD after five-epoch smoothing. FORTE attains the lowest final mean loss on Tox21, SIDER, and ToxCast, whereas Uniform KD is slightly lower on MUV.

Additional routing compositions. Figure 9 completes Figure 5 with the mean and seed-wise routing masses on MUV, Tox21, SIDER, and ToxCast.

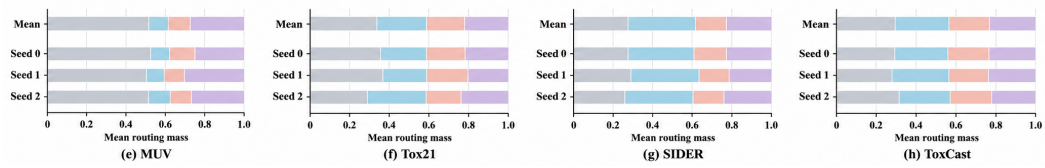

Figure 9: (e)–(h) Mean and seed-wise routing masses assigned to Null, IEM, EDG, and SMI on MUV, Tox21, SIDER, and ToxCast, respectively, completing Figure 5. The null route receives a comparably large share on MUV, while the allocations are comparatively stable across seeds on SIDER and ToxCast.

C.8 Teacher Availability and Routing Across Classification Tasks

To examine whether the routing behavior extends across classification tasks, we compare teacher coverage and paired contributions before and after routing on all eight datasets using the GIN validation diagnostics with seeds 0, 1, and 2. Figure 10 separates availability from the empirical cumulative distributions of available teacher contributions, $q_{k,i}s_{k,i}$ before routing and $\pi_{k,i}q_{k,i}s_{k,i}$ after routing. We observe negative contributions for every teacher on every dataset before routing. After routing, these negative tails disappear, while positive contributions remain in all 24 teacher–dataset combinations. The jump at zero records suppressed contributions, whereas the remaining positive tail shows that FORTE does not simply turn off all auxiliary supervision. The extent of reweighting differs across datasets and teachers, consistent with conditioning teacher influence on the downstream task rather than assigning fixed weights.

Missing teachers on MUV. MUV has substantially lower teacher coverage: IEM and EDG are each available for only 23.9% of the diagnostic molecules, whereas SMI covers 100%. We exclude unavailable observations from the contribution distributions rather than replacing them with zeros. Among available MUV observations, 58.6%, 56.9%, and 64.0% of the IEM, EDG, and SMI contributions are negative before routing; all three negative fractions become zero after routing, while each teacher retains a positive tail. Thus, the routing mechanism remains operative when some modalities are missing. Separately, the downstream MUV results in Table 1 show that FORTE improves ROC-AUC from 74.2 to 78.7 with GIN, from 68.3 to 77.1 with MolCLR, and from 78.3 to 79.4 with Mole-BERT. These benchmark gains complement the availability diagnostic, although they do not isolate the causal effect of missing teachers.

Findings. Across all eight datasets, FORTE suppresses negative teacher contributions while retaining positive auxiliary signals. On MUV, this selective routing remains visible despite sparse IEM and EDG coverage, and the downstream benchmark results show gains over all three corresponding students.

C.9 Extended Analysis of Task-Conditioned Routing

To examine how FORTE controls teacher influence at the molecular level, we extend the routing analysis to all eight classification datasets using the GIN validation checkpoints for seeds 0, 1, and 2. The following figures track individual teacher contributions, their combined task alignment, and the strength of retained auxiliary supervision. All comparisons use matched observations from the same checkpoints.

Figure 11 compares the signed contribution $q_{k,i}s_{k,i}$ with its routed counterpart $\pi_{k,i}q_{k,i}s_{k,i}$, keeping the molecule order fixed within each seed. Here $q_{k,i}$ is the calibrated teacher-to-task gradient norm ratio from Eq. (7), $s_{k,i}$ is the teacher-to-task cosine similarity from Eq. (8), and $\pi_{k,i}$ is the final routing weight. Thus, $q_{k,i}s_{k,i}$ measures the teacher gradient's task-aligned component relative to the task-gradient norm, using the same teacher-first index order as the main text. We observe both positive and negative contributions before routing, with different patterns across molecules and teachers. After routing, negative entries are suppressed to zero, while a subset of positive entries remains. The retained pattern varies within each teacher row, showing that selection operates at the molecule level rather than excluding a teacher globally. Gray entries remain unavailable in both conditions and are distinct from the zero contributions produced by routing.

Figure 12 shows whether this teacher-level selection also resolves conflicts in the combined auxiliary signal. The horizontal coordinate is the mean signed contribution over available teachers under equal weighting, and the vertical coordinate is the sum of routing-weighted contributions for the same molecule and checkpoint. We observe negative equal-weight components on every dataset, whereas all recorded routed components are nonnegative. Within the negative horizontal region, some observations move to positive routed alignment and others lie at zero. Thus, FORTE can retain compatible contributions even when their equal-weight mixture opposes the task, while withholding auxiliary updates when no eligible contribution remains.

We next examine the retained auxiliary-to-main gradient norm ratio U in Figure 13, grouping observations by the number of available teachers with positive compatibility and evidence. All zero-eligible observations have $U = 0$, whereas the other groups retain nonzero auxiliary strength. The recorded ratios remain below one across all eight datasets. Their distributions overlap and do not increase uniformly with teacher count, indicating that the number of eligible teachers alone does not determine auxiliary influence. These are conditional distributions within the diagnostic runs, not a controlled teacher-count ablation.

Routing with incomplete teachers on MUV. MUV has the largest zero-eligible group (56.0%), while 34.9%, 6.0%, and 3.1% of observations have one, two, and three eligible teachers, respectively. Eligibility combines availability with compatibility and evidence, so these shares should not be read as missing-modality rates. Despite the sparse IEM and EDG coverage reported in Section C.8, the one-eligible group retains a median auxiliary-to-main norm ratio of 0.73. We therefore observe selective abstention alongside substantial auxiliary learning on the remaining observations, rather than a uniform collapse of distillation on MUV. The separate MUV benchmark gains reported above support the practical usefulness of FORTE, without attributing those gains to missing-teacher handling alone.

Findings. Across eight datasets, FORTE removes task-opposing contributions while retaining nonzero, task-aligned auxiliary updates. Its auxiliary influence depends on the available evidence rather than teacher count alone; on MUV, frequent abstention coexists with substantial retained supervision.

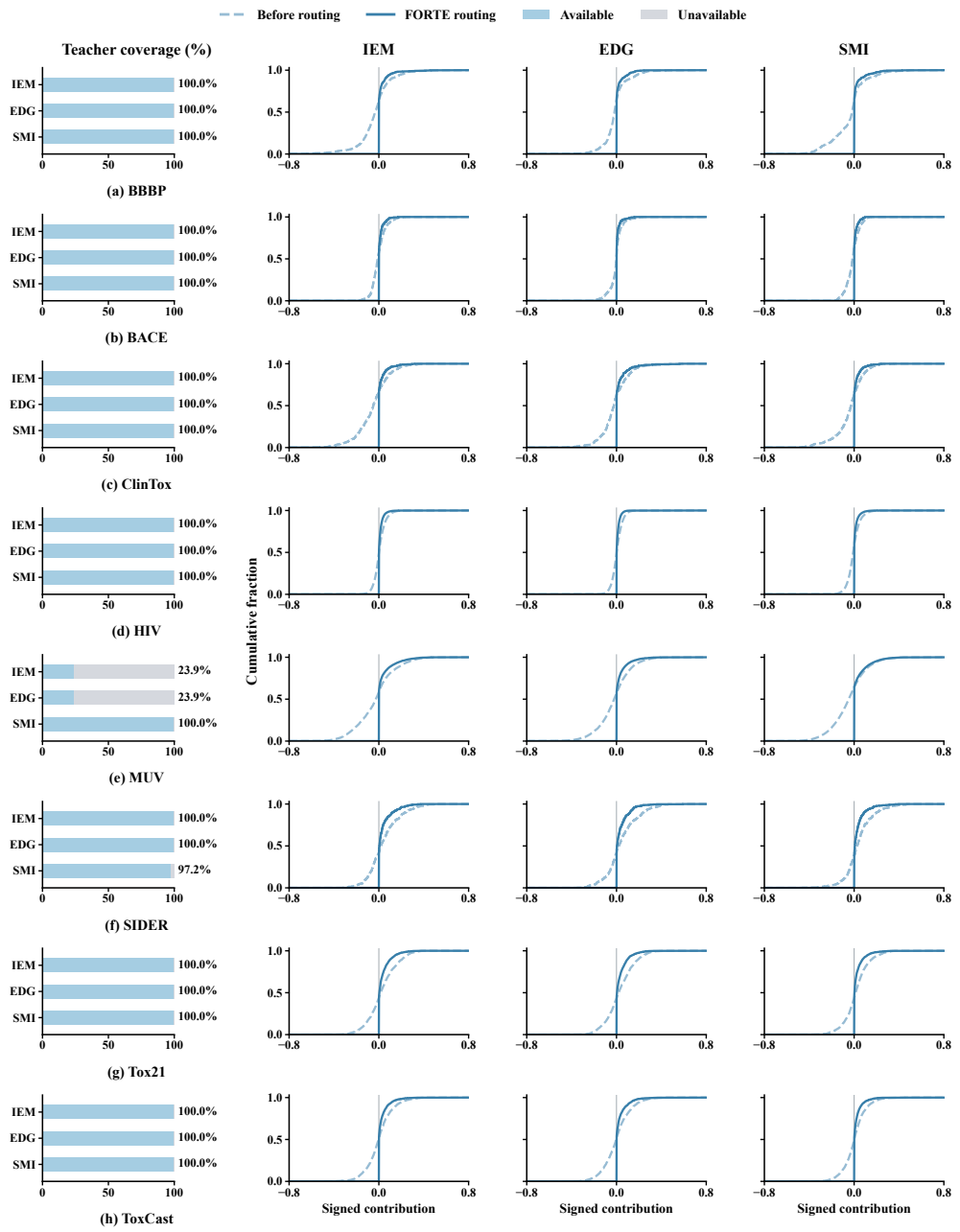

Figure 10: (a) BBBP, (b) BACE, (c) ClinTox, (d) HIV, (e) MUV, (f) SIDER, (g) Tox21, and (h) ToxCast: teacher coverage and empirical cumulative distributions of signed contributions before and after routing, pooled across three seeds and conditioned on teacher availability. FORTE removes negative contributions while retaining positive tails; gray coverage segments indicate unavailable teachers, most prominently on MUV.

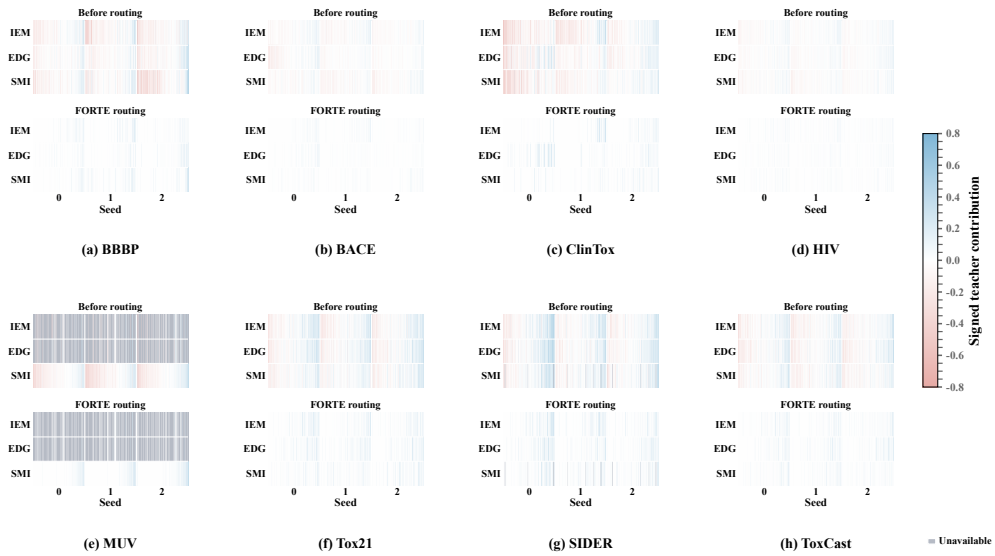

Figure 11: (a) BBBP, (b) BACE, (c) ClinTox, (d) HIV, (e) MUV, (f) Tox21, (g) SIDER, and (h) ToxCast: matched teacher contributions before and after routing, with columns grouped by seed and a shared signed color scale. Negative contributions disappear after routing while positive contributions remain molecule-dependent; gray entries denote unavailable teachers rather than suppressed signals.

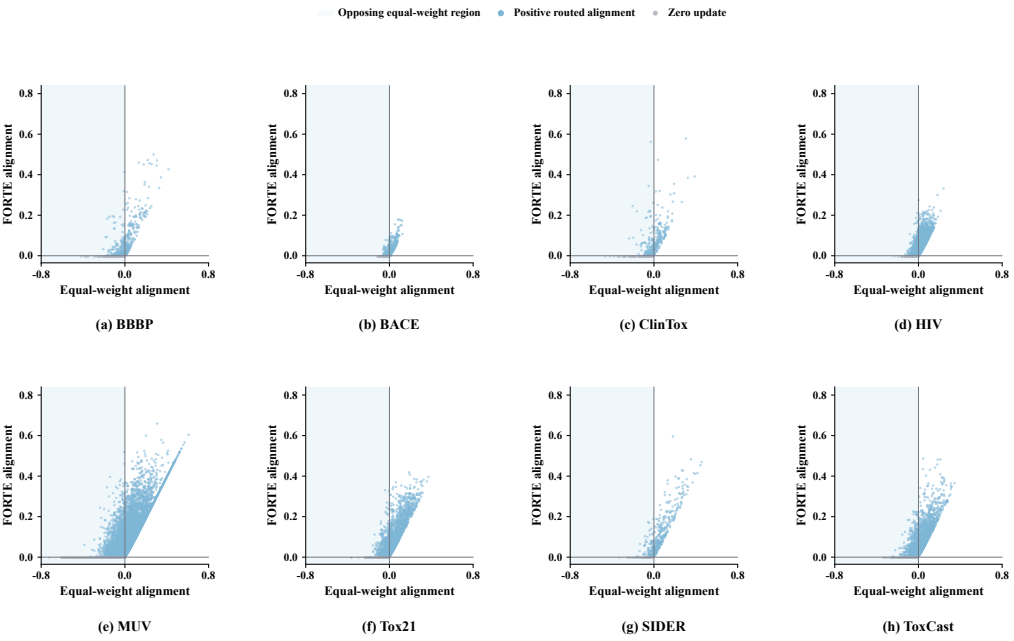

Figure 12: (a)–(h) show BBBP, BACE, ClinTox, HIV, MUV, Tox21, SIDER, and ToxCast, respectively, with equal-weight task-aligned components on the horizontal axis and FORTE components on the vertical axis. Observations in the shaded negative equal-weight region have positive or zero routed components, showing conflict suppression without uniformly discarding auxiliary supervision.

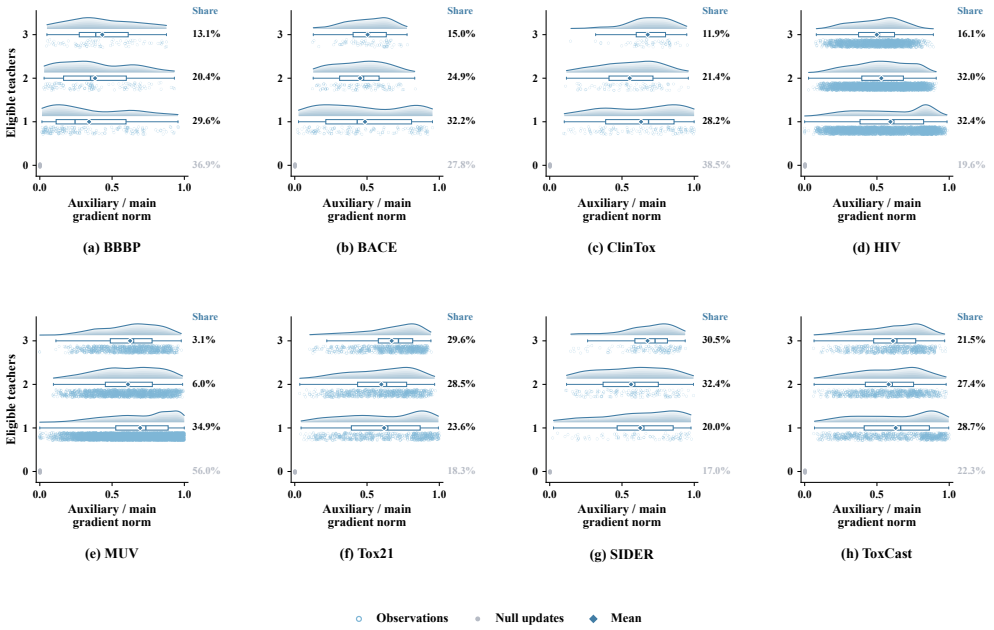

Figure 13: (a)–(h) show the auxiliary-to-main gradient norm ratio conditioned on eligible-teacher count for BBBP, BACE, ClinTox, HIV, MUV, Tox21, SIDER, and ToxCast, with group shares reported alongside the distributions. Zero-eligible observations receive no auxiliary update, while the remaining groups retain overlapping, nonzero strength distributions, including on MUV.